Przegląd TELEKOMUNIKACYJNY
WIADOMOŚCI TELEKOMUNIKACYJNE

4/2023

Dwumiesięcznik Stowarzyszenia Elektryków Polskich

ISSN 1230-3496, e-ISSN 2449-7487
Cena: 52 zł (w tym 8% VAT)

WYDAWNICTWO *SIGMA-NOT*

Szanowni Państwo, Koleżanki, Koledzy,

w imieniu swoim i całego Komitetu Programowego, a także Komitetu Organizacyjnego serdecznie zapraszam do udziału w Konferencji Radiokomunikacji i Teleinformatyki, która odbędzie się w dn. 20–22 września 2023 r. w Krakowie. Organizatorem Konferencji jest Instytut Telekomunikacji Akademii Górniczo-Hutniczej im. Stanisława Staszica w Krakowie.

Konferencja Radiokomunikacji i Teleinformatyki, choć w nowej odsłonie, nie jest wydarzeniem zupełnie nowym na polskiej mapie konferencji naukowych. Dobrze znane środowisku Krajowa Konferencja Radiokomunikacji, Radiofonii i Telewizji (KKRRiT) oraz Krajowe Sympozjum Telekomunikacji i Teleinformatyki (KSTiT) przez trzy ostatnie edycje były organizowane wspólnie pod szyldem Multikonferencji. W tym roku naszedł czas, by zintegrować je w jedno duże wydarzenie pod wspólną nazwą *Konferencja Radiokomunikacji i Teleinformatyki*. Każda z Konferencji bazowała dotąd na utrwalonym od lat bazowym programie. KKRRiT zrzeszała naukowców, wystawców i przedsiębiorców skupionych głównie wokół technik radiowych i multimedialnych, podczas gdy KSTiT najczęściej była miejscem spotkań i dyskusji dla specjalistów od telekomunikacji przewodowej i rozwiązań teleinformatycznych. Dziś oba te światy łączymy z nadzieją na szerszą współpracę środowiska telekomunikacyjnego w Polsce i jeszcze szybszy postęp naukowy i techniczny w tej ważnej dyscyplinie naukowej.

Telekomunikacja obejmuje szeroki zakres wiedzy i zagadnień, często łącząc się w wielu rozwiązaniach z osiągnięciami z zakresu elektroniki i informatyki. Stąd też w języku oficjalnym, jak tym bardziej potocznym na popularności zyskuje pojęcie Teleinformatyka. Jak sama nazwa wskazuje, pojęcie to nawiązuje wprost do wspólnego świata telekomunikacji i informatyki, który to świat jednakowoż stoi na mocnym fundamencie zaawansowanej elektroniki. Rozwiązania chmurowe, wirtualizacja, uczenie maszynowe, sztuczna inteligencja czy wszelkie zagadnienia związane z cyberbezpieczeństwem to najlepsze przykłady rozwiązań wprost związanych z teleinformatyką.

W obecnych trudnych czasach, spowodowanych w dużej mierze agresją Rosji na niepodległą Ukrainę, szczególnie ważne stają się zagadnienia związane z zapewnianiem bezpieczeństwa cyfrowego. Już od kilkunastu lat pojawiały się w świecie naukowym, ale także przekazach medialnych, informacje o coraz większym wpływie zagadnień związanych z cyberbezpieczeństwem na wyniki wyborów, kreowanie rzeczywistości czy coraz popularniejsze oszustwa internetowe. Dziś zapewnienie postępu naukowego i w jego wyniku zaawansowanych rozwiązań technicznych w zakresie cyberbezpieczeństwa staje się jeszcze ważniejsze i może okazać się kluczowe dla zachowania bezpieczeństwa i pokoju światowego.

Program Konferencji Radiokomunikacji i Teleinformatyki jest bardzo bogaty i zawiera szereg niezwykle ciekawych wystąpień, również w zakresie cyberbezpieczeństwa. Zaplanowaliśmy pięć referatów plenarnych, 16 sesji tematycznych, dwie sesje firmowe, sesje specjalne i posiedzenia komitetów. W trakcie Konferencji zaprezentowanych zostanie łącznie 96 referatów o różnej, bardzo aktualnej tematyce. Uczestnikami Konferencji będą naukowcy, doktoranci, studenci, a także przedstawiciele biznesu i sektora władz na szczeblu centralnym i samorządowym. To niezwykła okazja, by ciągu trzech dni poznać trendy i najnowsze rozwiązania z zakresu teleinformatyki.

Dziękuję wszystkim sponsorom, patronom honorowym, organizatorom konkursów dla młodych autorów, Stowarzyszeniu Inżynierów Telekomunikacji, URSI, a także fundatorom nagród, w tym firmie Sprint za podtrzymanie tradycji wręczenia Złotych Cyborgów.

Wszystkim autorom przyjętych artykułów serdecznie gratuluję i liczę na to, że wspólnie spędzimy niezapomniane chwile podczas KRiT 2023 w Krakowie!

Jerzy Domżał

Przewodniczący Komitetu Programowego

4/2023

DWUMIESIĘCZNIK STOWARZYSZENIA
ELEKTRYKÓW POLSKICH
WYDAWANY PRZY WSPÓŁPRACY
KOMITETU
ELEKTRONIKI I TELEKOMUNIKACJI
POLSKIEJ AKADEMII NAUK

Przegląd TELEKOMUNIKACYJNY

TELE–RADIO–ELEKTRONIKA–INFORMATYKA

ROK ZAŁOŻENIA 1928 • ROCZNIK XCVI • ISSN 1230-3496

ZESPÓŁ REDAKCYJNY

dr hab. inż. Janusz Dudczyk, prof. WAT
– redaktor naczelny

dr inż. Bogdan Zbierzchowski – z-ca redaktora naczelnego

Stały współpracownik: mgr inż. Cezary Rudnicki

Redaktorzy tematyczni: dr inż. Piotr Grzejszczak, dr hab. inż. Marek Suproniuk, prof. WAT, dr hab. inż. Marcin Wesołowski

Bożena Lachowicz – sekretarz redakcji
tel. (22) 827 38 79; 604 446 028
e-mail: **elektronika@red.pl.pl**

Adres redakcji:
ul. Czackiego 3/5, 00-043 Warszawa
www.przegladtelekomunikacyjny.pl
przeg.tel@sigma-not.pl

Zamówienia na reklamę przyjmuje Redakcja
lub Dział Reklamy i Marketingu,
ul. Ratuszowa 11, 03-450 Warszawa
tel. 22 827 43 65, e-mail: **reklama@sigma-not.pl**

Zakład Poligrafii i Kolportażu
Wydawnictwa SIGMA-NOT
ul. ks. J. Popiełuszki 19/21, 01-595 Warszawa,
tel. 22 840 35 89; tel. 22 840 59 49, 22 891 13 74,
e-mail: **prenumerata@sigma-not.pl**

RADA PROGRAMOWO-NAUKOWA

Prof. dr hab. inż. Józef Woźniak – przewodniczący
dr hab. Andrzej Bęben, prof. PW; Mgr inż. Jacek Cyrek, mgr inż. Paweł Biskupski; dr hab. inż. Jerzy Domżał, prof. AGH; dr inż. Andrzej Dulka; dr hab. inż. Sławomir Hausman, prof. PŁ; prof. dr hab. inż. Andrzej Jajszczyk; mgr inż. Stefan Kamiński; dr hab. inż. Jan Kelner, prof. WAT; dr hab. inż. Adrian Kliks, prof. PP; prof. dr hab. inż. Józef Modelski; dr hab. inż. Marek Natkaniec, prof. AGH; prof. dr hab. Stanisław Piątek; dr hab. inż. Jacek Rak, prof. PG; dr inż. Łukasz Rybak; dr hab. inż. Kajetana Snopek, prof. PW; dr hab. inż. Kamil Staniec, prof. PWr; prof. dr hab. inż. Jacek Stefański; dr hab. inż. Krzysztof Szczypiorski, prof. PW; dr hab. inż. Krzysztof Wesołowski, prof. PP; prof. dr hab. Tadeusz Więckowski; dr inż. Andrzej Wilk

Publikowane artykuły naukowe są recenzowane przez samodzielnych pracowników nauki

Redakcja nie ponosi odpowiedzialności za treść ogłoszeń. Zastrzega sobie prawo do skracania i adiustacji nadesłanych materiałów.

Indeks 35722 Nakład do 500 egz (w tym wersja elektroniczna)

Skład: Studio DTP Wydawnictwa SIGMA-NOT, Aleksandra Olszewska

WYDAWNICTWO SIGMA-NOT

03-450 Warszawa, ul. Ratuszowa 11
tel.: 22 818-09-18, 22 818-98-32
www.sigma-not.pl
sekretariat@sigma-not.pl

TELECOMMUNICATION REVIEW – TELECOMMUNICATION NEWS
Tele-Radio-Electronics, Information Technology

SPIS TREŚCI • CONTENTS

SPIS REFERATÓW ZAMIESZCZONYCH NA PORTALU INFORMACJI TECHNICZNEJ SIGMA-NOT 3

REFERATY PLENARNE

ZASTOSOWANIE UCZENIA MASZYNOWEGO W LOKALNYCH SIECIACH BEZPRZEWODOWYCH
Application of machine learning in wireless local area networks
S. SZOTT 7

PRZEŻYWALNOŚĆ ELASTYCZNYCH SIECI OPTYCZNYCH ZE ZWIELOKROTNIENIEM PRZESTRZENNYM
Survivability of elastic optical networks with space division mutliplexing
R. GOŚCIEŃ 15

ULTRA-SZYBKA TRANSMISJA ŚWIATŁOWODOWA NA POTRZEBY CENTRÓW DANYCH
Ultra-highs speed optical fibre transmission in data centers
J. TURKIEWICZ 24

METAMATERIAŁY, KONFIGUROWALNE MATRYCE ANTENOWE I KOMUNIKACJA HOLOGRAFICZNA
– wstępna analiza nowej koncepcji bezprzewodowej transmisji danych
Metamaterials, reconfigurable antenna arrays and holographic communication – preliminary analysis of a new concept of wireless data transmission
A. KLIKS 29

WYZWANIA SPOŁECZNE, EKONOMICZNE I TECHNICZNE WDROŻENIA TRANSPORTU AUTONOMICZNEGO
Social, economic and technical challenges of autonomous transport implementation
S.. KOSIELIŃSKI 40

Autor za publikację artykułu w czasopiśmie naukowym *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne* otrzymuje **20 punktów** zgodnie z komunikatem Ministra Edukacji i Nauki z dnia 1 grudnia 2021 r. w sprawie wykazu czasopism naukowych i recenzowanych materiałów z konferencji międzynarodowych.

RADA PROGRAMOWO-NAUKOWA

Prof. dr hab. inż. Józef Woźniak – PRZEWODNICZĄCY
Dr hab. Andrzej Bęben , prof. PW – Politechnika Warszawska
Mgr inż. Jacek Cyrek – WB Electronics
Mgr inż. Paweł Biskupski – Systemics PAB
Dr hab. inż. Jerzy Domżał, prof. AGH – Akademia Górniczo-Hutnicza w Krakowie
Dr inż. Andrzej Dulka – PIIT
Dr hab. inż. Sławomir Hausman, prof. PŁ – Politechnika Łódzka
Prof. dr hab. inż. Andrzej Jajszczyk – Akademia Górniczo-Hutnicza w Krakowie
Mgr inż. Stefan Kamiński – KIGEiT
Dr hab. inż. Jan Kelner, prof. WAT – Wojskowa Akademia Techniczna
Dr hab. inż. Adrian Kliks, prof. PP – Politechnika Poznańska
Prof. dr hab. inż. Józef Modelski – Politechnika Warszawska
Dr hab. inż. Marek Natkaniec, prof. AGH – Akademia Górniczo-Hutnicza w Krakowie
Prof. dr hab. Stanisław Piątek – Uniwersytet Warszawski
Dr hab. inż. Jacek Rak, prof. PG – Politechnika Gdańska
Dr inż. Łukasz Rybak – Wojskowa Akademia Techniczna
Dr hab. inż. Kajetana Snopek, prof. PW – Politechnika Warszawska
Dr hab. inż. Kamil Staniec, prof. PWr – Politechnika Wrocławska
Prof. dr hab. inż. Jacek Stefański – Politechnika Gdańska
Dr hab. inż. Krzysztof Szczypiorski, prof. PW – Politechnika Warszawska
Dr hab. inż. Krzysztof Wesołowski, prof. PP – Politechnika Poznańska
Prof. dr hab. Tadeusz Więckowski – Politechnika Wrocławska
Dr inż. Andrzej Wilk

WYDAWNICTWO *SIGMA-NOT*

WARIANTY PRENUMERATY 2023

Prenumerata papierowa – czasopismo tylko w wersji papierowej, za jego dostarczenie doliczamy opłatę roczną;
Prenumerata cyfrowa – czasopismo wyłącznie w wersji cyfrowej, dostępne na Portalu Informacji Technicznej www.sigma-not.pl, prenumerator otrzyma indywidualny link dostępu do logowania na Portalu;
PRENUMERATA PAKIET PLUS – czasopismo w wersji papierowej (bez opłaty za dostarczanie prasy) oraz cyfrowej, a także dostęp do archiwum zaprenumerowanego tytułu na Portalu Informacji Technicznej www.sigma-not.pl wraz z indywidualnym linkiem do logowania.

Cena jednego egzemplarza – 52 zł

CENY PRENUMERATY ROCZNEJ:

- prenumerata papierowa: 300 zł brutto + (koszty wysyłki 18 zł);
- prenumerata cyfrowa: 300 zł brutto;
- prenumerata w wersji Pakiet PLUS (papierowa + cyfrowa + archiwum) 450,- zł.

PORTAL INFORMACJI TECHNICZNEJ www.sigma-not.pl – to największa polska internetowa baza czasopism i artykułów technicznych, wyposażona w szybką i skuteczną wyszukiwarkę tematyczną, umożliwiającą dostęp on-line do ponad 129 000 publikacji z różnych dziedzin.
Dla Państwa wygody logowanie do Portalu Informacji Technicznej stało się łatwiejsze. Portal zyskał nową, przejrzystą szatę graficzną, pojawiły się nowe funkcje ułatwiające zakup pojedynczych artykułów i całych zeszytów. Każdy prenumerator wersji cyfrowej oraz prenumeraty Pakietu PLUS otrzyma indywidulany link z kodem dostępu do zamówionego czasopisma.

Cena brutto zawiera 8% VAT na czasopisma w wersji papierowej oraz cyfrowej. W przypadku zmiany przez ustawodawcę stawki VAT na czasopisma i w konsekwencji zmiany ceny brutto, prenumerator zobowiązany jest do dopłaty różnicy.
Dla prenumeratorów zagranicznych obowiązuje cena według kursu waluty NBP (z dnia bezpośrednio poprzedzającego datę wystawienia faktury) plus koszty wysyłki.

PRENUMERATĘ ZAMAWIAMY
22 840 35 89
prenumerata@sigma-not.pl
www.sigma-not.pl

LISTA RECENZENTÓW

Prof. dr hab. Adam Abramowicz
Prof. dr hab. Marek Amanowicz
Prof. dr hab. Andrzej Dobrowolski
Prof. dr hab. Piotr Gajewski
Prof. dr hab. Marcin Iwanowski
Dr hab. Jacek Izydorczyk
Prof. dr hab. Tomasz Kacprzak
Prof. dr hab. Andrzej Karwowski
Prof. dr hab. Ryszard Katulski
Prof. dr hab. Bogdan Kwolek
Prof. dr hab. Roman Kubacki
Prof. dr hab. Lidia Łukasiak
Prof. dr hab. Stanisław Osowski
Dr hab. inż. Andrzej Paszkiewicz
Prof. dr hab. Krzysztof Perlicki
Prof. dr hab. Grzegorz Różański
Prof. dr hab. Jacek Stefański
Dr hab. Zenon Szczepaniak
Prof. dr hab. Maciej Walkowiak
Prof. dr hab. Krzysztof Wesołowski
Prof. dr hab. Ryszard Zieliński

Niniejszy zeszyt (4/2023) naszego dwumiesięcznika jest związany z KONFERENCJĄ RADIOKOMUNIKACJI i TELEINFORMATYKI 2023. Konferencja powstała w wyniku połączenia Krajowej Konferencji Radiokomunikacji, Radiofonii i Telewizji (KKRRiT) oraz Krajowego Sympozjum Telekomunikacji i Teleinformatyki (KSTiT). Organizatorem tegorocznej KRiT jest Instytut Telekomunikacji Akademii Górniczo-Hutniczej im. Stanisława Staszica w Krakowie.

Zeszyt zawiera całość materiałów konferencyjnych. Referaty plenarne są drukowane, a pełne teksty wszystkich referatów są zamieszczone na Portalu Informacji Technicznej SIGMA-NOT z nieodpłatnym dostępem: https://sigma-not.pl/zeszyt-7248-przeglad-telekomunikacyjny-2023-4.html

Stanowią one integralną część zeszytu. Wszystkie artykuły i referaty są recenzowane. Podstawowe informacje na temat konferencji znajdą Państwo w artykule wstępnym (II str. okładki), przygotowanym przez dr hab. inż. Jerzego Domżała, prof. AGH – przewodniczącego konferencji KRiT 2023.

LISTA RECENZENTÓW

Amanowicz Marek, Ambroziak Sławomir Jerzy, Batalla Jordi Mongay, Bęben Andrzej, Biernacka Edyta, Blok Marek, Bogucka Hanna, Boryło Piotr, Chołda Piotr, Chydziński Andrzej, Czachórski Tadeusz, Danilewicz Grzegorz, Domański Marek, Domżał Jerzy, Głąbowski Mariusz, Goścień Róża, Grzybkowski Maciej J., Hausman Sławomir, Janowski Lucjan, Juszka Dawid, Kabaciński Wojciech, Kaczmarek Sylwester, Katulski Ryszard, Kelner Jan, Kliks Adrian, Klinkowski Mirosław, Konorski Jerzy, Kosek-Szot Katarzyna, Kostek Bożena, Krenz Rafał, Kucharski Andrzej, Kula Sławomir, Kułakowski Paweł, Leszczuk Mikołaj, Łopatka Jerzy, Matiolański Andrzej, Mazurczyk Wojciech, Michalak Jarosław, Misiurewicz Jacek, Modelski Józef, Natkaniec Marek, Niemiec Marcin, Pach Andrzej R., Pacyna Piotr, Papir Zdzisław, Pastuszak Grzegorz, Perlicki Krzysztof, Piotrowski Zbigniew, Remlein Piotr, Romaniuk Ryszard, Różański Grzegorz, Różycki Marek, Sadowski Jarosław, Sęga Wiktor, Siergiejczyk Mirosław, Słobodzian Piotr, Słowik Stanisław, Snopek Kajetana, Stasiak Maciek, Stefański Jacek, Suchański Marek, Sułek Wojciech, Szczepański Paweł, Szott Szymon, Tarasiuk Halina, Wajda Krzysztof, Walkowiak Krzysztof, Wesołowski Krzysztof, Wójcik Robert, Wojtasiak Wojciech, Woźniak Józef, Yashchyshyn Yevhen, Zieliński Ryszard J., Zieliński Tomasz P., Zwierzykowski Piotr

Referaty sesji tematycznych zamieszczone na Portalu Informacji Technicznej SIGMA-NOT

ST01. APLIKACJE I USŁUGI

A. MATIOLAŃSKI
WSPÓŁCZESNE ARCHITEKTURY SYSTEMÓW OPARTYCH O MIKROSERWISY 49

K. ROMAN, J. KOŁAKOWSKI
SYSTEM DO MONITOROWANIA NARAŻENIA NA WIBRACJE. 53

S. KACZMAREK, M. KRASOŃ
EMULACJA ŚRODOWISKA DLA ZASTOSOWANIA PROTOKOŁU IN-BAND NETWORK TELEMETRY 57

D. ZMYSŁOWSKI, J. KELNER
SKORELOWANIE METRYK QoS DLA USŁUGI PRZEGLĄDANIA STRON INTERNETOWYCH W SIECIACH MOBILNYCH LTE. 61

S. UDDIN, M. LESZCZUK, M. GREGA
QUANTIZATION PARAMETER-BASED VIDEO QUALITY ASSESSMENT IN MPEG CODECS LEVERAGING VMAF 65

ST02. ASPEKTY RADIOWE SIECI 5G I 6G

K. WESOŁOWSKI
NOWA METODA WYKRYWANIA INTENCJONALNYCH ZAKŁÓCEŃ W TRANSMISJI OFDM 69

P. KARPOVICH, T. P. ZIELIŃSKI
VEHICLE DETECTION IN 6G SYSTEMS WITH OTFS MODULATION 73

M. HOFFMANN
BADANIE SIECI MASSIVE MIMO O ARCHITEKTURZE ZORIENTOWANEJ NA UŻYTKOWNIKA 77

M. KUSTRA, B. GROCHOWINA, K. KOSMOWSKI
NIEZAKŁÓCONA PRYWATNA SIEĆ 5G NA POTRZEBY SYTUACJI KRYZYSOWEJ. 81

S. JANJI, P. SROKA, A. KLIKS
WYKORZYSTANIE REKONFIGUROWALNYCH INTELIGENTNYCH MATRYC ANTENOWYCH W ŁĄCZU DOSYŁOWYM SIECI 5G/6G WYKORZYSTUJĄCEJ BEZZAŁOGOWE STATKI POWIETRZNE 85

ST03. ASPEKTY SIECI 5G I 6G

M. ZIĘBA, P. BORYŁO
CLOSED-ENABLED DEPLOYMENT OF 5G CORE NETWORK 89

Y. ALI, W. SUŁEK
MEASUREMET ANALYSIS OF ERROR RATE FOR BELIEF'S CLIPPING AND QUANTIZATION IN POLAR CODES 93

A. SAMORZEWSKI, A. KLIKS
SYSTEMY KOMÓRKOWE 5G WSPIERANE PRZEZ BSP, OZE ORAZ IPR 97

E. BIERNACKA, M. KLINKOWSKI
ANALIZA DYNAMICZNEGO WYBORU WEZŁA PRZETWARZAJĄCEGO Z ALOKACJĄ PRZEPŁYWÓW W PAKIETOWYCH SIECIACH XHAUL . . . 101

T. KOSIŁO, J. MARSKI , K. RADECKI
BADANIE TERMINALA 5G DO ZASTOSOWAŃ IoT 105

ST04. BEZPIECZEŃSTWO SIECI

W. SZCZEPANIK, M. NIEMIEC
UCZENIE MASZYNOWE W SIECIACH KWANTOWEJ DYSTRYBUCJI KLUCZA . . . 109

K. SOBOLAK, P. CHOŁDA
UCZENIE MASZYNOWE Z OCHRONĄ PRYWATNOŚCI DANYCH . . 113

A. VEJAR, P. CHOŁDA
MACHINE LEARNING BASED NETWORK ANOMALY DETECTION WITH PROTECTION AGAINST ADVERSARIAL ATTACKS 117

P. BEZEG, P. PACYNA
ANALIZA WSKAŹNIKA Φ POD KĄTEM PRZYDATNOŚCVI JAKO MIARY W OCENIE ZŁOŻONOŚCI SZYFROWANIA I JAKOŚCI SZYFROGRAMU W METODZIE S-TECH. 121

J. WOJTUŃ, K. BEDNARZ, K. MAŚLANKA
WYSOCE BEZPIECZNA ŁĄCZNOŚĆ SPECJALNA NA BAZIE NOWOCZESNYCH RADIOSTACJI WOJSKOWYCH 125

P. RYDLICHOWSKI, A. KLIKS
INFRASTRUKTURY KOMUNIKACJI KWANTOWEJ – SYNERGIE ORAZ INTEGRACJA Z SYSTEMAMI TRANSMISJI WZORCOWYCH SYGNAŁÓW CZASU I CZĘSTOTLIWOŚCI 129

ST05. KONKURS NA NAJLEPSZĄ PRACĘ DOKTORSKĄ

K. ABRATKIEWICZ
DEKOMPOZYCJA IMPULSÓW RADAROWYCH ZA POMOCĄ TRIANGULACJI ZER SPEKTROGRAMU 133

J. BIENIASZ, K. SZCZYPIORSKI
ROZPROSZONE METODY UKRYWANIA INFORMACJI W SIECIACH . 137

B. BOSSY
EFEKTYWNA ENERGETYCZNIE ALOKACJA ZASOBÓW W SYSTEMACH OFDM ZE ŚWIADOMOŚCIĄ OBLICZENIOWĄ 143

Ł. KUŁACZ
WYKORZYSTANIE INFORMACJI KONTEKSTOWEJ DO POPRAWY EFEKTYWNOŚCI WIDMOWEJ W SYSTEMACH RADIOWYCH 147

R. SMOLARZ
MIKROFALOWE SPRZĘGACZE KIERUNKOWE O POLEPSZONYCH PARAMETRACH I ICH ZASTOSOWANIE W MONOLITYCZNYCH MIKROFALOWYCH UKŁADACH SCALONYCH. 151

J. SOBOLEWSKI
ZASTOSOWANIE ELEKTROD GRAFENOWYCH W PRZEŁĄCZNIKACH NA PASMO FAL MILIMETROWYCH WYKONANYCH W STRUKTURACH AlGaN/GaN. 155

ST06. LOKALNE SIECI BEZPRZEWODOWE WLAN

G. TECA, M. NATKANIEC
STEGOBACKOFF: CREATING A COVERT CHANNEL USING BACKOFF PROCEDURE IN IEEE 802.11 NETWORKS 159

S. STRYCZEK, M. NATKANIEC
NEGATYWNY WPŁYW NISKICH POZIOMÓW MOCY NADAWANIA NA WYDAJNOŚĆ PRACY SIECI BEZPRZEWODOWYCH STANDARDU IEEE 802.11AX. 164

M. GWIAZDOWICZ, M. NATKANIEC
ANALIZA MOŻLIWOŚCI ŚWIADCZENIA USŁUG Z ODPOWIEDNIĄ JAKOŚCIĄ W WIELOŁĄCZOWYCH SIECIACH STANDARDU IEEE 802.11BE. 169

M. WOJNAR, W. CIĘŻOBKA
ZASTOSOWANIE ALGORYTMÓW WIELORĘKICH BANDYTÓW DO WYBORU SZYBKOŚCI TRANSMISJI W SIECIACH IEEE 802.11AX 173

M. NATKANIEC, M. KASZUBA
ANALIZA DZIAŁANIA MECHANIZMÓW POTWIERDZEŃ RAMEK DLA TRYBU MU OFDMA SIECI STANDARDU IEEE 802.11AX 177

A. JAMROŻY, K. KOSEK-SZOTT
WDROŻENIE KONTROLERA SDN DO ZARZĄDZANIA PUNKTAMI DOSTĘPOWYMI STANDARDU IEEE 802.11 182

ST07. METODY SZTUCZNEJ INTELIGENCJI

Y. SALMI, H. BOGUCKA
ATTACK AGAINST COMMUNICATION AND COMPUTING TASK CLASSIFICATION AND MITIGATION TECHNIOQUES. 186

O. FIEDOT, M. BLOK
DEKODOWANIE KODÓW ITEROWANYCH Z UŻYCIEM SIECI NEURONOWEJ 190

K. ZUCHNIAK, E. KOSK, P. JAGLARZ,
SMOKEFINDER – WSPOMAGANY ALGORYTMAMI SZTUCZNEJ INTELIGENCJI SYSTEM WIZYJNEJ DETEKCJI DYMU. 194

M. WASILEWSKA, H. BOGUCKA
ANALIZA WPŁYWU ATAKÓW ZATRUWAJĄCYCH NA DETEKCJĘ ZASOBÓW RADIOWYCH BAZUJĄCĄ NA UCZENIU FEDERACYJNYM. . 198

D. DULAS, K. WALKOWIAK
AI-ASSISTED DIMENSIONING OF 5G NETWORK SLICES – REVIEW AND PERSPECTIVES 204

S. ZARĘBSKI, M. JOKIEJ, M. SZCZUKIEWICZ, K. RUSEK, P. CHOŁDA
AN UNSUPERVISED MACHINE LEARNING APPROACH FOR REGRESSION TESTING QUEUE OPTIMISATION IN 5G BASE STATIONS 208

ST08. METODY SZTUCZNEJ INTELIGENCJI W SYSTEMACH RADIOWYCH

K.J. MANDECKI, M. KOŁAKOWSKI
WYKORZYSTANIE INFORMACJI NA TEMAT AKTYWNOŚCI UŻYTKOWNIKA DO ZWIĘKSZENIA DOKŁADNOŚCI W ULTRASZEROKOPASMOWYM SYSTEMIE LOKALIZACYJNYM . . 212

P. GAJEWSKI, J. ŁOPATKA, R. CHĘCIŃSKI, A. KASZUBA-CHĘCIŃSKA
ZARZĄDZANIE POLITYKĄ W MOBILNEJ SIECI RADIA KOGNITYWNEGO Z DYNAMICZNYM DOSTĘPEM DO WIDMA 216

S. BETKA, M. BLOK
ZASTOSOWANIE SIECI NEURONOWYCH DO DETEKCJI IMPULSÓW O ZNANYM KSZTAŁCIE W OBECNOŚCI SILNEGO SZUMU I TRENDU 220

R. KRENZ, M. MICHALSKI, I. ZAINUTDINOV
SYSTEM ŁĄCZNOŚCI DLA SATELITÓW CUBESAT – ARCHITEKTURA I PROTOKOŁY. 224

K. ZUCHNIAK, E. KOSK, P. JAGLARZ
FIREFINDER – SIEĆ DETEKTORÓW ŚRODOWISKOWYCH, WYKORZYSTUJĄCYCH AI I NB-IOT UMOŻLIWIAJĄCYCH WCZESNĄ DETEKCJĘ POŻARÓW W LASACH 228

A. OLEJNICZAK, K.K. CWALINA, J. SADOWSKI, J. STEFAŃSKI
ANALIZA SKUTECZNOŚCI DETEKCJI SYGNAŁÓW GMSK W KANALE Z PROPAGACJĄ WIELODROGOWĄ PRZY UŻYCIU METOD GŁĘBOKIEGO UCZENIA 232

ST09. OPTOTELEKOMUNIKACJA I SYSTEMY RADIOLOKALIZACJI

J. STEFAŃSKI, J. SADOWSKI
SYSTEM RADIOLOKALIZACYJNY OPARTY NA FARMIE TURBIN WIATROWYCH 236

R. KAWECKI, S. HAUSMAN, P. KORBEL
ZASTOSOWANIE KRIGINGU MAP RSSI DO LOKALIZACJI WEWNĄTRZBUDYNKOWEJ METODĄ ODCISKÓW PALCÓW WSPOMAGANEJ FILTRACJĄ CZĄSTECZKOWĄ. 240

J. SADOWSKI, J. STEFAŃSKI
WSKAZYWANIE PUNKTU STARTOWEGO DLA ITERACYJNYCH ALGORYTMÓW OBLICZANIA POŁOŻENIA Z UŻYCIEM SIECI NEURONOWEJ 244

W. ZĄBEK, P. JURKIEWICZ, R. WÓJCIK, J. DOMŻAŁ
OPTYMALIZACJA ENERGETYCZNA SZKIELETU SIECI Z UŻYCIEM BYPASSÓW OPTYCZNYCH. 248

S. KULA, W. KOWALSKA
ODPORNOŚĆ SIECI xPON NA DZIAŁNIE NIEPOSŁUSZNEJ JEDNOSTKI ONU 252

ST10. POMIARY W TELEKOMUNIKACJI

K. BEDNARZ, J. WOJTUŃ
STABILNOŚĆ CZĘSTOTLIWOŚCI NISKOBUŻDZETOWYCH PLATFORM SDR W PROCEDURACH LOKALIZACJI OPARTYCH O METODY DOPPLEROWSKIE 256

A. ZIĘBA, J. SADOWSKI, M. KOLLAR, K. TATARCZYK
PRECYZYJNOŚĆ POMIARU PRZEPŁYWNOŚĆI IP W SIECI LTE Z FUNKCJĄ DUAL CONNECTIVITY. 260

P. JURKIEWICZ, M. RZEPKA, P. JAGLARZ,
J. DOMŻAŁ, R. WÓJCIK
DYNAMICZNE DOSTOSOWYWANIE WAG ŁĄCZY DOSTĘPOWYCH W MULTI-HOMINGU. 265

M. SUMIŁA
OTOCZENIE RADIOWE LINII KOLEJOWYCH W KONTEKŚCIE ZAKŁÓCEŃ ŁĄCZNOŚCI W SYSTEMIE GSM-R 269

M. ZAMŁYŃSKI
METODY POMIARU PARAMETRÓW ANTEN STOSOWANYCH W KARTACH RFID 273

Z. ZAKRZEWSKI
INTEGRACJA O-RAN ORAZ XPON NA WARSTWIE OPTYCZNYCH ZASOBÓW ODN 277

ST11. PROPAGACJA FAL RADIOWYCH

Ł. KUŁACZ, A. KLIKS
WYKORZYSTANIE REKONFIGUROWALNYCH MATRYC ANTENOWYCH WRAZ Z INFORMACJĄ KONTEKSTOWĄ 281

J. KAWECKI, Ł. JANUSZKIEWICZ, P. DI BARBA
MODEL DWUWYMIAROWY DO SYMULACJI GĘSTOŚCI MOCY W OBSZARZE GŁOWY I OCZU NA CZĘSTOTLIWOŚCI 26 GHZ 285

M. KRYK, K. MALON, J. KELNER
MAPY ŚRODOWISKA RADIOWEGO BAZUJĄCE NA PEM W SIECIACH FANET 289

J. WOJTUŃ, C. ZIÓŁKOWSKI, J. KELNER
UPROSZCZONE MODELE CHARAKTERYSTYK ANTENOWYCH W ZASTOSOWANIU DO OCENY INTERFERENCJI MIĘDZYWIĄZKOWYCH 293

B. ŚLESICKI, A. ŚLESICKA, A. KAWALEC, M. ŻMIGRODZKA
ANALIZA PORÓWNWAWCZA METOD ESTYMACJI MACIERZY KOWARIANCJI ZAKŁÓCEŃ W PRZESTRZENNO-CZASOWYM ADAPTACYJNYM PRZETWARZANIU SYGNAŁÓW 297

B. ULJASZ, C. ZIÓŁKOWSKI
ANALIZA PORÓWNAWCZA MODELI PROFILU GĘSTOŚCI ŁADUNKÓW JONOSFERY WYKORZYSTYWANYCH DO PREDYKCJI SYGNAŁÓW W KANAŁACH KF 301

ST12. TECHNIKA ANTENOWA

P. OLEKSY, Ł. JANUSZKIEWICZ
ANTENA ENDOSKOPU KAPSUŁKOWEGO PRZEZNACZONA DO PRACY W PASMIE MICS 400 MHZ 305

L. NOWOSIELSKI, B. DUDZIŃSKI, M. NOWOSIELSKI, A.M. ŚLUBOWSKA
KONCEPCJA ZDALNEGO STEROWANIA BLOKU WZMACNIACZY ANTENOWYCH SYSTEMU ANTENOWEGO AM 524 W KOMORACH SAC 309

T. WALCZYNA, Z. PIOTROWSKI
IMPLEMENTACJA METODY UKRYWANIA DANYCH W OBRAZACH Z WYKORZYSTANIEM SIECI GENERATYWNYCH 313

R. PRZESMYCKI, M. BUGAJ
WPŁYW CIAŁA LUDZKIEGO NA PARAMETRY ELEKTRYCZNE ANTENY TEKSTRONICZNEJ PRZEZNACZONEJ DO PRACY W SYSTEMIE 5G 317

R. PRZESMYCKI, M. BUGAJ
POMIAR CHARAKTERYSTYK PROMIENIOWANIA ANTEN Z WYKORZYSTANIEM IMPULSÓW ELEKTROMAGNETYCZNYCH HPM 321

ST13. TECHNIKI MULTIMEDIALNE

T. KONASZYŃSKI, D. JUSZKA, M. LESZCZUK
ANALIZA WPŁYWU MIKRO-STRUKTURY PREZENTACJI TREŚCI NA SUBIEKTYWNĄ OCENĘ JAKOŚCI WIDEO ORAZ JEJ STABILNOŚĆ 325

A. VEJAR, T. KONAZYŃSKI, D. JUSZKA, M. LESZCZUK
UNSUPERVISED CATEGORY LEARNING IN PERCEIVED VIDEO QUALITY 329

aiseemo

ANALITYKI WIDEO OPARTE NA SZTUCZNEJ INTELIGENCJI
DO AUTOMATYZACJI MONITORINGU WIZYJNEGO

WYKRYWANIE ZDARZEŃ W CZASIE RZECZYWISTYM

PRZYKŁADOWE ROZWIĄZANIA:

- Detekcja osób i ich cech – maseczki medyczne, kaski, odzież ochronna
- Detekcja pojazdów i ich cech – rozpoznawanie tablic rejestracyjnych, kategorie
- Detekcja osób w paśmie termograficznym z kamer hybrydowych lub multispektralnych
- Poprawa jakości obrazu wideo za pomocą technologii Super Resolution
- Detekcja uszkodzeń, wad i wandalizmu kamer
- Detekcja ognia i dymu
- Detekcja broni

KORZYŚCI:

- Analityki wideo dostosowane do specyficznych potrzeb
- Elastyczność i skalowalność - usługa w modelach licencyjnym lub SaaS
- Większa efektywność monitoringu bez kosztownej modernizacji systemu
- Łatwe wdrożenie w infrastrukturze Klienta lub w chmurze
- Możliwość integracji dodatkowych analityk w przyszłości

YOU CAN SEE MORE...

Aiseemo sp. z o.o.
ul. Czarnowiejska 36 lok. C5/017 30-054 Kraków
kontakt@aiseemo.com www.aiseemo.com

A. DUTTA, M. LESZCZUK, D. JUSZKA, M. GREGA
SUBJECTIVE QUALITY ASSESSMENT OF VIDEO SUMMARISATION ALGORITHMS: A CROWDSOURCING APPROACH 333

P. DUSZEJKO, Z. PIOTROWSKI
ALGORYTM UKRYWANIA DANYCH W OBRAZIE OPARTY NA KLASYFIKACJI DO WYKRYWANIA I ROZPOZNAWANIA UKRYTYCH INFORMACJI 337

M. BISTROŃ, Z. PIOTROWSKI
ZWIĘKSZANIE POTENCJAŁU SZTUCZNEJ INTELIGENCJI GENERATYWNEJ DZIĘKI PROMPT ENGINEERING 341

ST14. TRANSMISJA RADIOWA W SYSTEMACH INTERNETU RZECZY

J. MAGIERA
SELEKTYWNE ZAKŁÓCANIE TRANSMISJI OFDM W INTERFEJSIE RADIOWYM NB-IoT 345

A. STRZODA
ZASTOSOWANIE HEURYSTYKI DO WYBORU LOKALIZACJI WĘZŁÓW POŚREDNICZĄCYCH W SIECIACH LP WAN 349

W. HRYCENKO, M. TWAROWSKI, V. DJAJA-JOŚKO
INFRASTRUKTURA SYSTEMU DO MONITOROWANIA AKTYWNOŚCI OSÓB STARSZYCH 353

J. JĘDRYS, J. KOŁAKOWSKI
SYSTEM DO BADANIA AKTYWNOŚCI OSÓB STARSZYCH Z WYKORZYSTANIEM PARAMETRÓW ŚRODOWISKOWYCH 357

K. DUSZCZYK, J. KOŁAKOWSKI
SYSTEM DO BADANIA CHODU OSÓB STARSZYCH 361

ST15. ZAGADNIENIA WARSTWY FIZYCZNEJ W NOWOCZESNYCH SIECIACH RADIOKOMUNIKACYJNYCH

K. HAUSTEIN, M. BLOK
KOD FONTANNOWY Z PRZYRASTAJĄCĄ LICZBĄ SYMBOLI ŹRÓDŁOWYCH 365

J. GAJEWSKI, J. ŁOPATKA
WYZNACZENIE WARTOŚCI MOCY SYGNAŁU MASKUJĄCEGO SIEĆ PRZED NAMIERZENIEM PRZY WYKORZYSTANIU ALGORYTMU MUSIC 369

M. MAZURO
SENSOR MONITORINGU WIDMA ZINTEGROWANY Z BEZZAŁOGOWYM STATKIEM POWIETRZNYM 373

W. KARCZ
WYZNACZANIE NATĘŻENIA POLA, GĘSTOŚCI MOCY I SAR W OTOCZENIU RADIOKOMUNIKACYJNYCH STACJI BAZOWYCH. 377

J. MARTYNA
ADAPTACYJNE KSZTAŁTOWANIE CHARAKTERYSTYKI PROMIENIOWANIA W MILIMETROWYCH SIECIACH 5G 382

S. SABEEH, K. WESOŁOWSKI
ADAPTACJA DYNAMICZNEGO ROZDZIAŁU ZASOBÓW W SYSTEMIE KOMUNIKACJI V2X W TRYBIE 1 W SIECIACH 5G NEW RADIO. 387

ST16. ZARZĄDZANIE SIECIAMI

W. TEKIELA, J. DOMŻAŁ
AUTOMATYZOWANIE PROCESÓW ZARZĄDZANIA ZMIANAMI W SIECIACH 391

C.A. CŁAPA, A. LASOŃ, V. KANITKAR, P. CHOŁDA
PRACTICAL MODELS OF INTERCONNECTIONS BETWEEN CONTENT DELIVERY NETWORKS AND INTERNET SERVICE PROVIDERS 395

T. WICHARY, J. MONGAY BATALLA
ZARZĄDZANIE WIELOWARSTWOWYMI ROZPROSZONYMI ZASOBAMI W SIECIACH 5G-ADVANCED I 6G Z ZACHOWANIEM BEZPIECZEŃSTWA I PRYWATNOŚCI DANYCH 399

B. KĄDZIOŁKA, P. JURKIEWICZ
REDUKCJA WIELKOŚCI TABLICY PRZEPŁYWÓW DZIĘKI DETEKCJI DUŻYCH PRZEPŁYWÓW ZA POMOCĄ UCZENIA MASZYNOWEGO 403

A. MARDAUS, J. DOMŻAŁ
PORÓWNANIE OTWARTOŹRÓDŁOWYCH IMPLEMENTACJI STEROWNIKÓW SDN: OPENDAYLIGHT, OPEN NETWORK OPERATING SYSTEM, FLOODLIGHT 407

K. RADAMSKI, R. WÓJCIK
MECHANIZM ADAPTACYJNEGO ZARZĄDZANIA CZASAMI ŻYCIA PRZEPŁYWÓW W SIECIACH SDN. 411

SYSTEMICS PAB
Network Quality Experts

Rubinowy sponsor konferencji KRiT

Rozwiązania dla prywatnych sieci 5G
Niezawodne rozwiązania do symulacji GNSS

Zastosowanie 5G
w zarządzaniu
kryzysowym i przemyśle

DOI: 10.15199/59.2023.4.1

Zastosowanie uczenia maszynowego w lokalnych sieciach bezprzewodowych

Application of machine learning in wireless local area networks

STRESZCZENIE: Pomocą w rozwiązaniu wyzwań stojących przed kolejnymi generacjami lokalnych sieci bezprzewodowych mogą okazać się metody oparte na uczeniu maszynowym (*machine learning*, ML). Celem niniejszego referatu jest przybliżenie tematyki zastosowania ML do poprawy wydajności sieci IEEE 802.11. Przedstawiono rodzaje uczenia maszynowego oraz przykłady zastosowań w sieciach Wi-Fi. Opisano działalność grupy IEEE 802.11 AIML TIG oraz wyznaczono przyszłe kierunki badań.
SŁOWA KLUCZOWE: IEEE 802.11, lokalne sieci bezprzewodowe, uczenie maszynowe, Wi-Fi.

ABSTRACT: Methods based on machine learning (ML) can prove to be helpful in addressing the challenges faced by future generations of wireless local area networks. The aim of this paper is to provide an overview of the application of ML to improve the performance of IEEE 802.11 networks. Various types of machine learning and examples of their applications in Wi-Fi networks are presented. The activities of the IEEE 802.11 AIML TIG are described and future research directions are outlined.
KEYWORDS: IEEE 802.11, machine learning, wireless local area networks, Wi-Fi.

dr hab. inż. SZYMON SZOTT, prof. AGH ,
AGH Akademia Górniczo-Hutnicza, Kraków,
szymon.szott@agh.edu.pl

Lokalne sieci bezprzewodowe działające według standardu IEEE 802.11 (komercyjnie znane jako Wi-Fi) odgrywają istotną rolę we współczesnej telekomunikacji. Według najnowszych statystyk, w roku 2023 na świecie będzie działało prawie 20 miliardów urządzeń tego standardu [27]. Jest wiele przyczyn rynkowego sukcesu Wi-Fi a jednym z nich jest ciągły rozwój standardu, który obecnie liczy ponad 4000 stron [8]. Ciągłe wprowadzanie nowych funkcjonalności powoduje, że kolejne generacje Wi-Fi są coraz bardziej skomplikowane i posiadają coraz większą liczbę parametrów konfiguracyjnych. Przykładowo, urządzenia 802.11be (Wi-Fi 7) działające w paśmie 6 GHz, które mogą dobrać szerokość kanału na 46 sposobów (w Europie) lub nawet 112 sposobów (w USA i innych krajach) [9]. Dodatkowo, pojawiają się nowe aplikacje korzystające z bezprzewodowej transmisji Wi-Fi (m.in. wirtualna/rozszerzona rzeczywistość, masowa komunikacja między maszynami, ang. *machine-to-machine*) o ścisłych wymaganiach w zakresie niskich opóźnień i wysokiej niezawodności. Gdy dodatkowo uwzględnimy cechy charakterystyczne wdrożeń sieci Wi-Fi (zmienne środowisko radiowe, gęste i nieskoordynowane rozmieszczenie punktów dostępowych, mobilność użytkowników, rozproszone zarządzanie, itd.), wyłoni się główne wyzwanie dla sieci IEEE 802.11 w przyszłości, którym jest dostosowanie działania coraz bardziej złożonych urządzeń do wymagań aplikacji, z uwzględnieniem zmiennych warunków środowiskowych. Znalezienie odpowiednich konfiguracji w ogromnej przestrzeni możliwych opcji za pomocą tradycyjnych algorytmów może okazać się

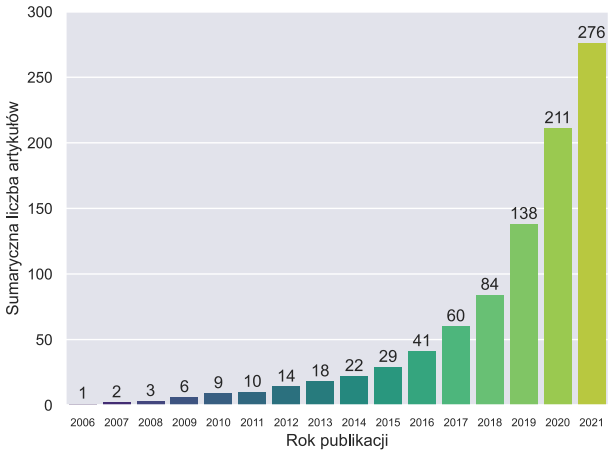
zbyt czasochłonne lub wymagające znaczących zasobów obliczeniowych.

W rozwiązaniu tego problemu pomocne może okazać się uczenie maszynowe (ang. *machine learning*, ML), rodzaj sztucznej inteligencji, w którym „algorytmy mogą uczyć się na podstawie danych treningowych bez konieczności jawnego ich zaprogramowania" [24]. Zalety ML do poprawy wydajności Wi-Fi dostrzegli naukowcy (*rys. 1*). Praca przeglądowa [20] obejmuje ponad 250 artykułów naukowych z tej tematyki a kolejne publikacje wciąż się ukazują (także na niniejszej konferencji [28]). Potencjał ML dostrzegła także grupa robocza IEEE 802.11, powołując specjalną podgrupę tematyczną do zbadania wpływu pojawiających się rozwiązań ML na przyszły kształt standardu [12]. Jest to zatem temat ważny i aktualny.

Celem referatu jest przybliżenie tematyki zastosowania uczenia maszynowego w lokalnych sieciach bezprzewodowych. Ograniczono się do zagadnień poprawy wydajności sieci Wi-Fi, gdyż uczenie maszynowe może być także zastosowane do poprawy bezpieczeństwa (np. wykrywania ataków) czy wykrywania aktywności (ang. *Wi-Fi sensing*). W rozdziale 2 pokrótce przedstawione są rodzaje ML oraz przykłady zastosowań w sieciach Wi-Fi. Rozdział 3 zawiera opis działalności standaryzacyjnej w tej tematyce. W rozdziale 4 wymieniono przyszłe kierunki badań, a rozdział 5 stanowi podsumowanie.

POPRAWA WYDAJNOŚCI

Praca przeglądowa [20] wskazuje, że uczenie maszynowe ma szerokie zastosowanie w obszarze poprawy wydajności sieci Wi-Fi: od warstwy fizycznej do metod zarządzania takimi sieciami. W *tabeli 1* przedstawiono główne obszary zastosowań uczenia maszynowego do poprawy wydajności sieci IEEE 802.11, które skupiają się na zwiększeniu przepływności (rozumianej jako chwilowe zwiększenie transmitowanej informacji w jednostce czasu). W dalszej części zostaną przedstawione trzy główne kategorie ML: uczenie nadzorowane (ang. *supervised learning*, SL), uczenie nienadzorowane (ang. *unsupervised learning*, UL) oraz uczenie ze wzmocnieniem (ang. *reinforced learning*, RL). Dla każdej kategorii podano definicję, typowe metody oraz możliwość zastosowania w lokalnych sieciach bezprzewodowych. Warto zwrócić uwagę, że w każdej z trzech głównych kategorii stosowane są sieci neuronowe.



» *Rys. 1. Sumaryczna liczba artykułów naukowych opublikowanych w obszarze poprawy wydajności Wi-Fi za pomocą uczenia maszynowego, stan z roku 2022 [20].*
» *Fig. 1. The total number of research articles published in the area of improving Wi-Fi performance using machine learning, as of 2022 [20].*

Uczenie nadzorowane

W uczeniu nadzorowanym (SL), zastosowany model uczy się funkcji odwzorowującej dane wejściowe na dane wyjściowe dzięki danym treningowym zawierającym zarówno wartości wejściowe, jak i wyjściowe [3]. Typowymi metodami SL są m.in. regresje, drzewa decyzyjne oraz maszyny wektorów nośnych (ang. support vector machines, SVM). SL używa się głównie do klasyfikacji (tj. przypisania danych do pewnych kategorii) lub do predykcji wartości.

Metody SL są często stosowane w literaturze do poprawy wydajności sieci IEEE 802.11, np. do wyboru punktu dostępowego, do którego stacja powinna się przyłączyć. W pracy [4] zaproponowano metodę, w której stacja wybiera ten punkt dostępowy (ang. *access point*, AP), który z największym prawdopodobieństwem zapewni najwyższą przepływność na podstawie wcześniejszych wyników. W tej metodzie wielowarstwowa sieć neuronowa uczy się korelacji między obserwacjami (np. siłą odebranego sygnału radiowego czy opóźnienia ramek *beacon*) a uzyskaną przepływnością. Metodę zweryfikowano eksperymentalnie wykazując jej skuteczność w porównaniu do tradycyjnych strategii wyboru AP.

Uczenie nienadzorowane

W uczeniu nienadzorowanym (UL), zastosowany model uczy się funkcji odwzorowującej dane wejściowe na dane

» *Tab. 1. Obszary zastosowań ML w sieciach IEEE 802.11*
» *Tabl. 1. Areas of applying ML in IEEE 802.11 networks*

Obszar	Typowy cel ML	Typowe metody	Zysk z ML
Dostęp do kanału radiowego	Wybór wartości okna współzawodnictwa	RL	Większa przepływność
Dostosowanie łącza	Wybór prędkości transmisji	RL	Większa przepływność
Warstwa fizyczna	Klasyfikacja sygnału, estymacja szumów	SL	Większa skuteczność klasyfikacji
Kształtowanie wiązki	Wybór parametrów (np. sektor)	SL	Wyższa przepływność, krótszy czas poszukiwań
Tryb *multi-user*	Przypisanie zasobów użytkownikom	SL i RL	Większa przepływność
Spatial reuse	Wybór kanału, mocy nadawczej	RL	Większa przepływność, mniejsze opóźnienia

» *Tab. 2. Zastosowanie ML w scenariuszu airport hotspot*
» *Tabl. 2. Application of ML in the airport hotspot scenario*

	Bez ML	ML w AP	ML w stacjach
Konfiguracja parametrów	Statyczna	Uaktualniana na bieżąco w AP i rozsyłana do stacji	Uaktualniana na bieżąco w stacjach
Dodatkowa sygnalizacja	Nie	Nie	Tak
Pomiary	Nie dotyczy	W AP	W stacjach
Optymalizacja parametrów	Brak/minimalna	Na podstawie wiedzy dostępnej w AP	Na podstawie wiedzy dostępnej w stacjach
Wydajność	Dobra	Lepsza	Najlepsza

wyjściowe dzięki danym treningowym zawierającym wyłącznie dane wejściowe [3]. Celem takiego procesu jest znalezienie wzorców w danych wejściowych, np. poprzez stworzenie klastrów podobnych danych. Typowymi metodami UL są m.in. analiza składowych głównych (ang. *principal component analysis*, PCA) oraz algorytm k-średnich (ang. *k-means*). Metody UL są rzadko stosowane w literaturze do poprawy wydajności sieci IEEE 802.11. Mogą służyć np. do wykrywania interferencji od innych technik bezprzewodowych [16] lub wspomóc wybór trybu pracy w transmisjach wielo-zakresowych (ang. *multi-band*) [26].

Uczenie ze wzmocnieniem

W uczeniu ze wzmocnieniem (RL), zastosowany model uczy się funkcji odwzorowującej dane wejściowe na dane wyjściowe dzięki interakcji ze środowiskiem: agent obserwuje środowisko, wybiera akcję i otrzymuje nagrodę [19]. Typowymi metodami RL są m.in. algorytmy wielorękich bandytów (ang. *multi-armed bandits*, MAB) czy Q-learning. Ważnym aspektem działania metod RL jest „kompromis pomiędzy eksploracją (próbą zdobycia nowej wiedzy) oraz eksploatacją (wykorzystaniem zdobytej wiedzy do podejmowania decyzji)" [28].

RL dobrze nadaje się do zastosowania w środowiskach lokalnych sieci bezprzewodowych ponieważ agent (w stacji lub AP), wykonuje pewne akcje (np. dobiera wartości parametrów bądź moment transmisji) na podstawie obserwacji zmiennego środowiska radiowego celem zwiększenia nagrody (np. uzyskiwanej przepływności). Przykładem jest rozwiązanie opisane w [29], gdzie punkt dostępowy monitoruje prawdopodobieństwo kolizji i wykorzystuje uczenie ze wzmocnieniem, aby wybrać wartość okna współzawodnictwa (ang. *contention window*, CW) dla wszystkich podłączonych do niego stacji. Dzięki temu możliwe jest utrzymanie wysokiej przepływności (wyższej niż w przypadku wyboru CW według reguł 802.11), zwłaszcza w przypadku zwiększającej się liczby transmitujących stacji.

ASPEKTY STANDARYZACYJNE

Wiedząc, że zastosowanie uczenia maszynowego może poprawić wydajność sieci Wi-Fi, kluczowym jest zbadanie w jaki sposób standard IEEE 802.11 może wesprzeć te działania. Raport [2], prezentowany w grupie roboczej IEEE 802.11, omawia kilka podstawowych wymagań. Po pierwsze, architektura 802.11 powinna wspierać działanie agentów ML, tj. procesów realizujących konkretny algorytm ML. Otwartą kwestią pozostaje czy w ramach pojedynczego urządzenia powinien działać jeden zaawansowany agent czy wiele współpracujących, ale prostszych agentów (każdy wspierający pojedynczą funkcjonalność). Potrzebne są protokoły oraz interfejsy programistyczne do wymiany danych między agentami jak również do wymiany kodu samych agentów. Przykładowo, kontroler sieciowy mógłby przesyłać kod agentów i instalować go przez interfejs programistyczny w punktach dostępowych czy stacjach klienckich. Należy także zapewnić wsparcie dla typowej pętli kontrolnej często wymaganej w rozwiązaniach opartych na ML: obserwacja stanu sieci, podjęcie decyzji bądź predykcji przez agenta ML, wdrożenie wybranej akcji. To z kolei wymaga odpowiednich zasobów obliczeniowych (do uruchamiania agentów i trenowania modeli) oraz sieciowych (do wymiany danych, modeli czy agentów). Wreszcie, ostatnim wymaganiem jest zdefiniowanie odpowiednich scenariuszy ewaluacyjnych by można było potwierdzić zasadność stosowania uczenia maszynowego.

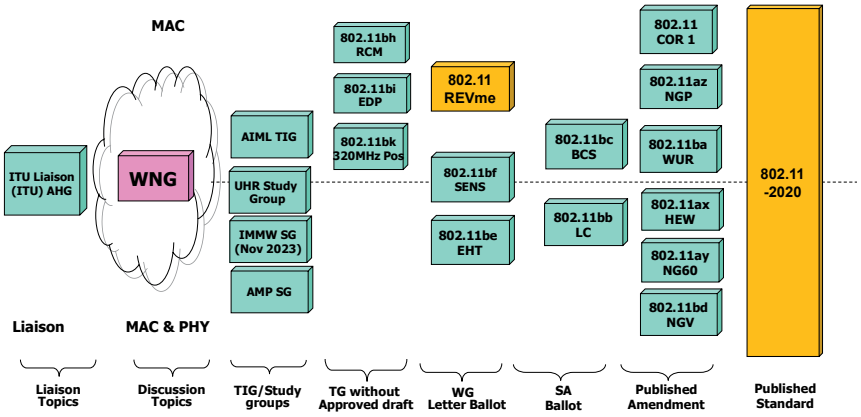
W *tabeli 2* przedstawiono przykładowy scenariusz zastosowania uczenia maszynowego w sieciach IEEE 802.11, który stanowi ilustrację wymienionych wymagań: zapewnienie dostępu do Internetu na lotnisku, ang. *airport hotspot*. To scenariusz charakteryzujący się gęstym rozmieszczeniem węzłów oraz zmieniającymi się warunkami a zatem taki, w którym jest miejsce na poprawę wydajności. Zestaw parametrów w tym scenariuszu obejmuje m.in. próg CCA (ang. *clear channel assessment*) oraz maksymalną moc transmisji. Bez zastosowania uczenia maszynowego punkt dostępowy ograniczony jest do rozsyłania statycznych (tj. uprzednio ustalonych) wartości parametrów w ramkach *beacon*. Jeśli wdrożymy uczenie maszynowe w punkcie dostępowym, to tak umiejscowiony agent może uaktualniać parametry w trakcie pracy sieci (dystrybuując je również przez ramki *beacon*) zapewniając tym samym lepszą wydajność niż dla standardowej sieci IEEE 802.11. Natomiast jeżeli punkt dostępowy może przesłać (podczas procesu przyłączania) wytrenowanego agenta do stacji, to taka stacja, uruchamiając agenta lokalnie, może na bieżąco aktualizować wybrane parametry i tym samym osiągać jeszcze lepszą wydajność niż w poprzednich konfiguracjach.

AIML TIG

Grupa robocza IEEE 802.11, wobec dużego zainteresowania zagadnieniami ML, postanowiła zbadać temat zastosowania uczenia maszynowego w lokalnych sieciach bezprzewodowych. W maju 2022 r. została powołana grupa tematyczna (ang. *topic interest group,* TIG) o nazwie *Artificial Intelligence Machine Learning* (AIML). Celem grupy jest „zbadanie przypadków użycia uczenia maszynowego dotyczących

» *Rys. 2. Grupy działające w ramach IEEE 802.11 [18]. Zazwyczaj nowe tematy badawcze wyłaniają się w ramach sesji Wireless Next Generation (WNG). Następnie powstają grupy badawcze (study groups), które definiują zakres dalszych prac w ramach grup zadaniowych (task groups), które z kolei przygotowują wersje robocze (draft) rozszerzeń. Te rozszerzenia podlegają procesowi zatwierdzania w formie balotażu (ballot). W końcu zostają opublikowane a z czasem wcielane są do głównego standardu 802.11.*

» *Fig. 2. Groups operating within IEEE 802.11 [18]. Typically, new research topics emerge during Wireless Next Generation (WNG) sessions. Then, study groups are formed, which define the scope of further work within task groups, which in turn prepare draft versions of standard extensions. These extensions are subject to a ballot approval process. Eventually they are published and over time incorporated into the main 802.11 standard.*



systemów i urządzeń IEEE 802.11, a także technicznej możliwości wdrożenia tych przypadków użycia" [12]. Podczas pierwszego spotkania AIML TIG, w lipcu 2022 r., zaprezentowałem stan badań nad zastosowaniem uczenia maszynowego do poprawy wydajności sieci IEEE 802.11 [21].

Na *rysunku 2* przedstawiono umiejscowienie AIML TIG wśród innych grup działających w ramach IEEE 802.11. Choć zazwyczaj grupy tego typu z czasem ewoluują w grupy zadaniowe (ang. *task group*, TG), które następnie tworzą rozszerzenie (ang. *amendment*) standardu 802.11, to wynikiem działania AIML TIG jest opracowanie raportu technicznego opisującego przypadki użycia (ang. use cases) uczenia maszynowego dla IEEE 802.11. Raport ten ma szanse wpłynąć na inne grupy działające w 802.11, a w szczególności na *Ultra High Reliability* (UHR) *Study Group* (SG), będącej prekursorem następnego głównego rozszerzenia standardu (802.11bn czyli Wi-Fi 8). Mandat grupy AIML TIG wygasa podczas zebrania plenarnego IEEE 802.11 we wrześniu 2023 r., zatem opracowywany raport techniczny jest w znacznej mierze gotowy [25]. Dlatego w następnym podrozdziale zostaną omówione główne przypadki użycia jakie wyłoniły się w toku dyskusji w AIML TIG.

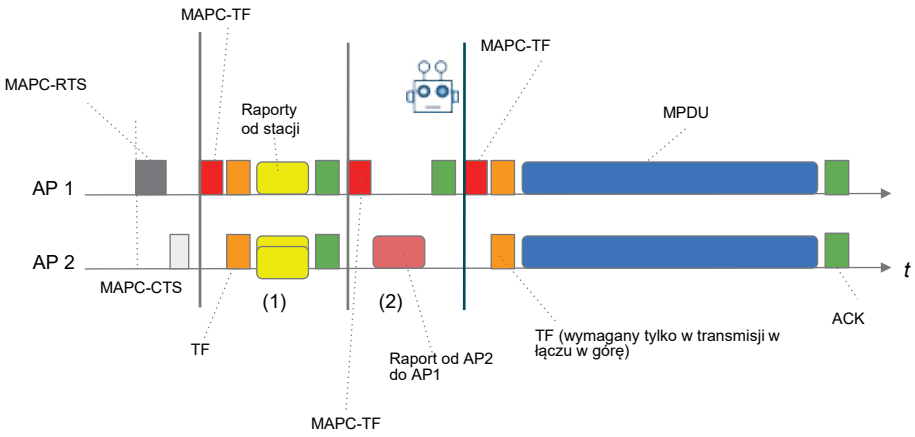
Zestandaryzowane przypadki użycia

Robocza wersja finalnego raportu grupy AIML TIG zawiera opis przypadków użycia dotyczących następujących funkcjonalności: kompresji informacji o stanie kanału, rozproszonego dostępu do kanału, wymiany modeli uczenia maszynowego między urządzeniami, oraz przełączania stacji przemieszczających się między AP [25]. Dodatkowo przygotowywanym przypadkiem użycia jest poprawa koordynacji wielu punktów dostępowych [22].

Pierwszy przypadek użycia dotyczy poprawy działania transmisji opartej na kształtowaniu wiązki (ang. *beamforming*), która wymaga zastosowania odpowiednich wag do sygnału transmitowanego przez układ wieloantenowy. Wagi te są wyliczane na podstawie informacji o stanie kanału (ang. *channel state information*, CSI). Raport z CSI jest wynikiem wymiany specjalnych ramek między nadawcą a odbiorcą w ramach procedury *channel sounding*. Ta wymiana ramek stanowi narzut sygnalizacyjny, który można zredukować za pomocą technik uczenia maszynowego. Przykładowo, na podstawie obserwacji, że CSI może być podobne dla stacji położonych blisko siebie, algorytm iFOR [6] stosuje metodę k-średnich, do kompresji CSI. Wyniki symulacji pokazują, że dla systemu z 8x2 MIMO, iFOR potrzebuje jedynie 8% dotychczas wymaganej liczby bitów do przesłania informacji zwrotnej co może zwiększyć przepływność nawet o 52% w stosunku do standardowego działania urządzeń IEEE 802.11ax. Raport [25] zwraca uwagę na następujące aspekty standaryzacyjne i wdrożeniowe: należy zachować kompatybilność wsteczną z dotychczasowymi metodami wymiany CSI, które jednocześnie mogą być używane jako dane treningowe; urządzenia używające kompresji CSI wspomaganej przez uczenie maszynowe muszą posiadać odpowiednie możliwości programistyczne i sprzętowe do korzystania z tych metod; zestandaryzowane metody powinny skracać czas potrzebny do wymiany informacji o CSI, jednocześnie minimalizując wszelki dodatkowy narzut (transmisyjny, obliczeniowy) wynikający z zastosowania uczenia maszynowego.

Drugi omawiany przypadek użycia dotyczy poprawy rozproszonej metody dostępu do kanału, która w 802.11 wywodzi się z protokołu wielodostępu CSMA/CA. W literaturze zaproponowano liczne modyfikacje tej metody [13]. W ostatnim czasie pojawiły się rozwiązania korzystające z metod uczenia się ze wzmocnieniem, które mogą nauczyć się optymalnych rozwiązań dla danych warunków sieciowych. Przykładem jest wspomniana wcześniej praca [29]. Raport [25] zwraca uwagę, że wdrażając zestandaryzowane rozwiązanie w tym zakresie należy zadbać o sprawiedliwe warunki dostępu do kanału dla wszystkich stacji, w tym stacji implementujących dotychczasowe rozwiązania (ang. *legacy*).

Kolejny przypadek użycia dotyczy wymiany modeli uczenia maszynowego między urządzeniami (np. punktem dostępowym a stacją kliencką). Przez modele ML należy rozumieć algorytmy implementujące pewien proces decyzyjny (np. wytrenowaną sieć neuronową). Zazwyczaj modele (jak również wymagane przez nie dane treningowe) wymieniane są jako dane warstwy aplikacji, co może być niewystarczające dla zastosowań o wysokich wymaganiach co do opóźnienia i niezawodności. Potrzebne są zatem efektywne sposoby wymiany modeli ML dla następujących trzech



» *Rys. 3 Przykładowa wymiana ramek w scenariuszu koordynacji transmisji przez wiele punktów dostępowych MAPC, opracowana na podstawie [22]. W fazie (1) zbierane są informacje od stacji, w fazie (2) – informacje są przekazywane tzw. udostępniającemu punktowi dostępowemu (tutaj: AP 1). Po fazie (2) agent ML podejmuje decyzje odnośnie dalszych transmisji. Przykład ma charakter ilustracyjny; dokładny sposób działania zostanie zestandaryzowany przez grupę UHR w 802.11bn.*

» *Fig. 3. An example of exchanging frames in the scenario of multi-AP transmission coordination, developed on the basis of [22]. In phase (1) information is collected from the stations, in phase (2) – information is transferred to the so-called sharing access point (here: AP 1). After phase (2), the ML agent makes decisions regarding further transmissions. The example is illustrative; the exact way of operation will be standardized by the UHR group in 802.11bn*

scenariuszy: uczenia scentralizowanego (stacje przesyłają dane treningowe do AP gdzie globalny model jest uaktualniany i rozsyłany do stacji), uczenia rozproszonego (zarówno stacje jak i AP wymieniają się danymi aby lokalnie trenować modele globalne), uczenie federacyjne (urządzenia trenują a następnie wymieniają się modelami lokalnymi; model globalny wyliczany jest w centralnym serwerze lub lokalnie). W tym celu raport [25] rekomenduje skorzystanie z rozgłoszeniowego charakteru transmisji w sieciach IEEE 802.11 wraz z powstającym rozszerzeniem 802.11bc (*enhanced broadcast services*, EBCS) [1].

Ostatni przypadek użycia rozważany w obecnej wersji raportu [25] dotyczy poprawy roamingu czyli przełączania się stacji klienckiej między punktami dostępowymi. Stacja podejmuje decyzję o przełączeniu na podstawie raportów otrzymanych od sąsiednich punktów dostępowych. W przypadku gęsto rozmieszczonych AP, uczenie maszynowe może przydzielić wagi (na podstawie wyuczonych wzorców przemieszczania się stacji) dla każdego AP w pobliżu, aby odpowiednio ustawić priorytety podczas uruchamiania procedury skanowania. Uczenie maszynowe może być również stosowane do predykcji poziomu siły otrzymanego sygnału (ang. *received signal strength indicator*, RSSI), przy którym należy rozpocząć procedurę skanowania. Poza odpowiednimi zasobami obliczeniowymi dostępnymi w stacji, takie rozwiązanie wymaga dodatkowej sygnalizacji w postaci rozszerzenia istniejących ramek zarządzających 802.11.

Dodatkowo rozważany przypadek użycia, opisany w [22], dotyczy funkcjonalności MAPC (ang. *multi-AP coordination*) czyli koordynacji jednoczesnych transmisji przez wiele punktów dostępowych w sytuacji nakładających się sieci. Jednoczesne transmisje mogą być realizowane np. przez odpowiedni podział zasobów w ramach OFDMA bądź inteligentne kształtowanie wiązki [23]. Funkcjonalność MAPC planowana jest w kolejnym rozszerzeniu standardu czyli 802.11bn, nad którą będzie pracowała grupa zadaniowa UHR. Na *rysunku 3* przedstawiono przykład wymiany ramek, która jest inicjowana przez tzw. udostępniający punkt dostępowy (AP 1) za pomocą wymiany ramek MAPC-RTS oraz MAPC-CTS między wszystkimi punktami dostępowymi biorącymi udział w nadchodzącej koordynowanej transmisji. Następnie, pierwsza faza wymiany (gromadzenie informacji) jest inicjowana przez udostępniający punkt dostępowy, który wysyła ramkę TF (ang. *trigger frame*) do pozostałych punktów dostępowych, które z kolei wysyłają ramki TF do swoich powiązanych stacji w celu zebrania wymaganych informacji. Stacje te odsyłają stosowne raporty. Należy zauważyć, że ta faza zbierania danych i raportowania może nie być konieczna, jeśli uczenie maszynowe może przewidywać raporty pomiarowe. W drugiej fazie, te raporty są przekazywane do udostępniającego punktu dostępowego, który korzysta z uczenia maszynowego do wyboru i konfiguracji strategii transmisji MAPC (OFDMA, kształtowania wiązki lub innej strategii), która zostanie użyta w ostatniej fazie transmisji. Oczekuje się, że wstępnie wytrenowany model zapewni albo niemal natychmiastową odpowiedź w celu niezwłocznego zaplanowania koordynowanej transmisji, albo zaplanuje koordynowane transmisje w niedalekiej przyszłości. Dotychczasowe wyniki z zastosowania ML do przydziału zasobów z użyciem OFDMA potwierdzają skuteczność tej metody [11].

Wyzwania standaryzacyjne

Standaryzacja rozwiązań używających uczenia maszynowego do poprawy wydajności sieci 802.11 obarczona jest kilkoma wyzwaniami. Przede wszystkim, zgodnie z koncepcją przyświecającą rozwojowi tego standardu, konieczne jest zapewnienie współistnienia z istniejącymi urządzeniami 802.11 w zakresie sygnalizacji oraz wydajności (tj. sprawiedliwości w dostępie do wspólnych zasobów). Innymi słowy, urządzenia legacy nie mogą być wykluczane z transmisji albo być gorzej traktowane.

Drugie wyzwanie jest związane z narzutem wprowadzanym przez używanie modeli ML. Z jednej strony, wymagają one czasu i mocy obliczeniowych na wytrenowanie a czasami mogą potrzebować dotrenowania w nowym środowisku. Z drugiej strony, modele potrzebują odpowiedniej ilości danych do podejmowania właściwych decyzji, co może skutkować wzrostem danych sygnalizacyjnych w sieci. Przykładowo, wadą rozwiązania z [29] jest konieczność

raportowania przez stacje do AP informacji o obserwowanym prawdopodobieństwie kolizji.

Trzecie wyzwanie związane jest z wymianą i uruchamianiem modeli na innych urządzeniach, np. gdy stacja otrzymuje od AP model ML do zarządzania daną funkcjonalnością 802.11. Jak wynika z dyskusji w AIML TIG, producenci są niechętni aby na ich urządzeniach był uruchamiany nieznany kod. Rozwiązanie tej kwestii bezpieczeństwa wymagałoby osobnych (tj. opracowanych poza standardem 802.11) metod weryfikacji kodu. Ten problem nie dotyczy rozwiązań niezależnych (ang. *stand-alone*) czyli takich, które działają lokalnie i nie potrzebują wymiany dodatkowej informacji z innymi urządzeniami. Przykładem takiej funkcjonalności są algorytmy wyboru szybkości transmisji (ang. *data rate selection*) [6]: standard 802.11 definiuje wyłącznie poziomy szybkości transmisji, a sposób ich wyboru pozostawia producentom sprzętu. Można się spodziewać, że producenci będą chcieli w niedalekiej przyszłości zastosować w tym obszarze uczenie maszynowe.

PRZYSZŁE KIERUNKI BADAŃ

Z analizy literatury badawczej oraz dyskusji standaryzacyjnych w ramach grupy roboczej IEEE 802.11 wyłaniają się omówione poniżej przyszłe kierunki badań w zakresie zastosowania uczenia maszynowego w sieciach Wi-Fi.

Rozwój IEEE 802.11

Standard IEEE 802.11 podlega ciągłemu rozwojowi a w kolejnych rozszerzeniach pojawiają się coraz bardziej skomplikowane funkcjonalności. Przykładowo, 802.11be (Wi-Fi 7) wprowadza połączenia korzystające jednocześnie z wielu pasm (ang. *multi-link operation*, MLO) a 802.11bn (Wi-Fi 8) najprawdopodobniej wprowadzi omawianą wcześniej skoordynowaną transmisję wielu punktów dostępowych (MAPC). Połączenie tych funkcjonalności z już wdrożonymi (łączeniem kanałów, OFDMA, kształtowaniem wiązki, itd.) skutkuje wysokim stopniem elastyczności w planowaniu transmisji w domenie czasu, przestrzeni i częstotliwości [20]. Metody uczenia maszynowego mogą być pomocne w znajdywaniu odpowiednich konfiguracji tych funkcjonalności, dostosowanych do panujących warunków radiowych i sieciowych.

Sieci Wi-Fi znajdują też nowe zastosowania, stąd rozwój standardu 802.11 w kierunku dwóch cech, które dotychczas nie były zapewniane w bezprzewodowych sieciach lokalnych, a które dotyczą usług wrażliwych na opóźnienie (ang. *time-sensitive*). Usługi takie wymagają zapewnienia precyzyjnego czasu transmisji, minimalnego opóźnienia oraz wysokiej niezawodności. Uczenie maszynowe może wspierać takie usługi np. planując dokładny czas ramki jak również przewidując jakość łącza.

Prace badawcze

Jest wiele możliwych kierunków rozwoju prac badawczych w tym obszarze. Analiza dotychczasowej literatury [20] wskazuje, że większość omawianych prac skupia się na optymalizacji pojedynczych funkcjonalności 802.11, np. na doborze wartości okna współzawodnictwa [29]. Tymczasem parametry konfiguracyjne powinny być optymalizowane wspólnie, aby osiągnąć lepszą wydajność w porównaniu do optymalizacji tych parametrów osobno. Przykładowo, jeśli chcemy poprawić wykorzystanie przestrzenne (ang. spatial reuse) należy łącznie konfigurować zarówno moc transmisji jak i próg detekcji nośnej. Zastosowanie uczenia maszynowego do łącznej optymalizacji kilku funkcjonalności 802.11 wydaje się obiecującym kierunkiem przyszłych badań.

W obszarze uczenia maszynowego znane jest stwierdzenie, że modele ML są tak dobre, jak dane, którymi zostały przeszkolone. Przyszłe prace badawcze mogą sięgnąć poza zwyczajowo używane dane z monitorowania stanu sieci. W grupie 802.11 intensywnie rozwijana jest, w ramach rozszerzeń 802.11mc oraz 802.11az, funkcjonalność lokalizacji urządzeń wewnątrz budynków na podstawie pomiarów czasowych. Znajomość położenia może służyć za daną wejściową do modeli uczenia maszynowego [5]. Pojawiły się też prace dotyczące kształtowania wiązki, gdzie model ML był trenowany za pomocą danych pochodzących z kamer wideo [15]. Można zatem spodziewać się więcej takich nieszablonowych rozwiązań.

Przyszłe prace badawcze powinny w szerszym zakresie dotyczyć eksperymentów z rzeczywistymi urządzeniami. Dotychczas większość badań, w tym ok. 80% artykułów omówionych w pracy przeglądowej [20], była ograniczona do badań symulacyjnych. Tymczasem badania eksperymentalne w środowisku rzeczywistym są kluczowe do weryfikacji proponowanych systemów. Szczególną uwagę należy zwrócić na możliwości sprzętowe urządzeń (zasoby obliczeniowe, czas pracy na baterii, itd.), które mają uruchamiać modele uczenia maszynowego. Przykładowo, popularne w sieciach IoT urządzenia wbudowane (ang. embedded devices) mogą nie wspierać najbardziej zaawansowanych metod głębokiego uczenia.

Stosowanie uczenia maszynowego

Pomimo popularności metod uczenia maszynowego (w wielu dyscyplinach inżynieryjno-technicznych), należy pamiętać, że niekoniecznie prowadzi ono do lepszych rozwiązań. Przykładowo, opracowany w pracy [10] model uczenia maszynowego ze wzmocnieniem nie okazał się znacząco lepszy od zaproponowanej wcześniej heurystyki. Dodatkowo wykazując przewagę modeli ML należy uwzględnić nie tylko poprawę wydajności, lecz również koszt wdrożenia takiego rozwiązania.

Tymczasem ograniczenia związane z dużą złożonością i znacznym czasem uczenia modeli ML mogą zostać przezwyciężone dzięki rozwijanym technikom uczenia rozproszonego. Należą do nich:

- uczenie hierarchiczne, które potrafi rozłożyć złożone problemy optymalizacyjne na wiele podproblemów [17],
- uczenie federacyjne, w którym modele są trenowane lokalnie a następnie agregowane w model globalny w centralnym węźle [30],
- uczenie transferowe, w którym model trenowany na jednym zadaniu jest używany ponownie do innego, powiązanego zadania [14].

Pozostaje kwestią sprawdzenia na ile te techniki uczenia rozproszonego można zastosować w lokalnych sieciach bezprzewodowych.

Uniwersalny interfejs

Patrząc w dalszą przyszłość rozwoju lokalnych sieci bezprzewodowych, można przez analogię do sieci sterowanych programowo rozważać koncepcję interfejsu radiowego sterowanego przez uczenie maszynowe (ang. *machine learning-driven radio interface*, MLDR), który byłby wolny od ograniczeń dotychczasowych technik bezprzewodowych. Zadaniem takiego interfejsu byłoby nauczenie się komunikowania, wybór i konfiguracja protokołów komunikacyjnych i tych funkcjonalności, które najlepiej pasują do konkretnego przypadku użycia i scenariusza, spełniając w ten sposób wymagania nowych aplikacji Internetowych i wydajnie wykorzystując dostępne zasoby radiowe. Innymi słowy, MLDR umożliwiłby wdrażanie „lokalnych sieci bezprzewodowych szytych na miarę". Podobna propozycja została przedstawiona w [7] dla sieci komórkowych.

PODSUMOWANIE

Popularność stosowania uczenia maszynowego w naukach inżynieryjno-technicznych nie ominęła lokalnych sieci bezprzewodowych. Odzwierciedleniem tego są zarówno liczne dotychczasowe i wciąż ukazujące się publikacje, jak również działalność grupy roboczej IEEE 802.11. Metody ML mogą, przy odpowiednim zastosowaniu, poprawić wydajność sieci Wi-Fi, w szczególności tam gdzie tradycyjne algorytmy nie prowadzą do optymalnych rozwiązań albo wymagają zbyt dużo zasobów obliczeniowych lub są zbyt czasochłonne. Należy przypuszczać, że zastosowanie uczenia maszynowego w sieciach IEEE 802.11 w najbliższych latach pozostanie nie tylko atrakcyjnym tematem badawczym, lecz także, że doczekamy się wdrożeń w urządzeniach, z których na co dzień korzystamy. ●

PODZIĘKOWANIE

Praca powstała w wyniku realizacji projektu badawczego o nr DEC-2020/39/I/ST7/01457 finansowanego ze środków Narodowego Centrum Nauki.

LITERATURA

[1] Ansley, Carol i inni. „802.11bc Functional Requirements Document." IEEE 802.11 September 2022 Interim Meeting. Dostęp 17 czerwca 2023. https://mentor.ieee.org/802.11/dcn/19/11-19-0151-05-00bc-802-11bc-functional-requirements-document.doc.

[2] Bellalta, Boris, Szymon Szott. 2022. „Wi-Fi Meets ML: Re-thinking Next Generation Wi-Fi Networks." IEEE 802.11 November 2022 Plenary Meeting. Dostęp 17 czerwca 2023. https://mentor.ieee.org/802.11/dcn/22/11-22-1784-01-0000-wi-fi-meets-ml-re-thinking-next-generation-wi-fi-networks-nov-tutorial.pptx.

[3] Berry, Michael W., Azlinah Mohamed, Bee Wah Yap. 2019. „Supervised and unsupervised learning for data science". Springer Nature.

[4] Bojovic, Biljana, Nicola Baldo, Jaume Nin-Guerrero, Paolo Dini. 2011. „A supervised learning approach to cognitive access point selection." 2011 IEEE GLOBECOM Workshops (GC Wkshps).

[5] Ciezobka, Wojciech, Maksymilian Wojnar, Katarzyna Kosek-Szott, Szymon Szott, Krzysztof Rusek. 2023. „FTMRate: Collision-Immune Distance-based Data Rate Selection for IEEE 802.11 Networks." 24th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM).

[6] Deshmukh, Mrugen, Mahmoud Kamel, Zinan Lin, Rui Yang, Hanqing Lou, Ismail Güvenç. 2022. „Intelligent Feedback Overhead Reduction (iFOR) in Wi-Fi 7 and Beyond." 2022 IEEE 95th Vehicular Technology Conference (VTC2022-Spring).

[7] Hoydis, Jakob, Fayçal Ait Aoudia, Alvaro Valcarce, Harish Viswanathan. 2021. „Toward a 6G AI-native air interface." IEEE Communications Magazine 59, no. 5:76-81.

[8] IEEE. 2020. „IEEE Standard for Information Technology--Telecommunications and Information Exchange between Systems - Local and Metropolitan Area Networks--Specific Requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications." IEEE Std 802.11-2020 (Revision of IEEE Std 802.11-2016).

[9] IEEE. 2022. „IEEE 802.11 Overview and Amendments under development". Dostęp 17 czerwca 2023. https://ieee802.org/11/IEEE%20802-11-Overview-and-Amendments-Under-Development.pptx.

[10] Kosek-Szott, Katarzyna, Szymon Szott, Falko Dressler. 2022. „Improving IEEE 802.11ax UORA Performance: Comparison of Reinforcement Learning and Heuristic Approaches." IEEE Access 10: 120285-120295.

[11] Kotagiri, Dheeraj, Koichi Nihei, Tansheng Li. 2021. „Distributed convolutional deep reinforcement learning based OFDMA MAC for 802.11ax." ICC 2021-IEEE International Conference on Communications.

[12] McCann, Stephen. 2022. „May 2022 Working Group Motions." IEEE 802.11 May 2022 Interim Meeting. Dostęp 17 czerwca 2023. https://mentor.ieee.org/802.11/dcn/22/11-22-0597-03-0000-may-2022-working-group-motions.pptx.

[13] Natkaniec, Marek, Katarzyna Kosek-Szott, Szymon Szott, Giuseppe Bianchi. 2012. „A survey of medium access mechanisms for providing QoS in ad-hoc networks." IEEE Communications Surveys & Tutorials 15, 2:592-620.

[14] Nguyen, Cong T., Nguyen Van Huynh, Nam H. Chu, Yuris Mulya Saputra, Dinh Thai Hoang, Diep N. Nguyen, Quoc-Viet Pham, Dusit Niyato, Eryk Dutkiewicz, Won-Joo Hwang. 2021. „Transfer learning for future wireless networks: A comprehensive survey." arXiv preprint arXiv:2102.07572.

[15] Nishio, Takayuki, Hironao Okamoto, Kota Nakashima, Yusuke Koda, Koji Yamamoto, Masahiro Morikura, Yusuke Asai, Ryo Miyatake. 2019. „Proactive received power prediction using machine learning and depth images for mmWave networks." IEEE Journal on Selected Areas in Communications 37, no. 11: 2413-2427.

[16] Olbrich, Michael, Anatolij Zubow, Sven Zehl, Adam Wolisz. 2017. „WiPlus: Towards LTE-U interference detection, assessment and mitigation in 802.11 networks." European Wireless 2017; 23th European Wireless Conference.

[17] Pateria, Shubham, Budhitama Subagdja, Ah-hwee Tan, and Chai Quek. 2021. „Hierarchical reinforcement learning: A comprehensive survey." ACM Computing Surveys (CSUR) 54, no. 5: 1-35.

[18] Stanley, Dorothy. 2023. „2023 May Working Group Chair Opening Report." IEEE 802.11 May 2023 Interim Meeting. Dostęp 17 czerwca 2023. https://mentor.ieee.org/802.11/dcn/23/11-23-0619-01-0000-2023-may-working-group-chair-opening-report.pptx.

[19] Sutton, Richard S., Andrew G. Barto. 2018. „Reinforcement learning: An introduction". MIT Press.
[20] Szott, Szymon, Katarzyna Kosek-Szott, Piotr Gawłowicz, Jorge Torres Gómez, Boris Bellalta, Anatolij Zubow, Falko Dressler. 2022. „Wi-Fi meets ML: A survey on improving IEEE 802.11 performance with machine learning." IEEE Communications Surveys & Tutorials 24, no. 3: 1843-1893.
[21] Szott, Szymon, Katarzyna Kosek-Szott, Piotr Gawłowicz, Jorge Torres Gómez, Boris Bellalta, Anatolij Zubow, Falko Dressler. 2022. „Applying ML to 802.11: Current Research and Emerging Use Cases." IEEE 802.11 July 2022 Plenary Meeting. Dostęp 17 czerwca 2023. https://mentor.ieee.org/802.11/dcn/22/11-22-0979-01-aiml-applying-ml-to-802-11-current-research-and-emerging-use-cases.pptx.
[22] Szott, Szymon, Katarzyna Kosek-Szott, Boris Bellalta. 2023. „Proposed IEEE 802.11 AIML TIG Technical Report Text for the Multi-AP Coordination Use Case". IEEE 802.11 July 2022 Plenary Meeting. Dostęp 17 czerwca 2023. https://mentor.ieee.org/802.11/dcn/23/11-23-0227-04-aiml-proposed-ieee-802-11-aiml-tig-technical-report-text-for-the-multi-ap-coordination-use-case.docx.
[23] Verma, Shikhar, Tiago Koketsu Rodrigues, Yuichi Kawamoto, Nei Kato. 2023. „A Survey on Multi-AP Coordination Approaches over Emerging WLANs: Future Directions and Open Challenges." arXiv preprint arXiv:2306.04164.
[24] Wang, Jingjing, Chunxiao Jiang, Haijun Zhang, Yong Ren, Kwang-Cheng Chen, and Lajos Hanzo. 2020. „Thirty years of machine learning: The road to Pareto-optimal wireless networks." IEEE Communications Surveys & Tutorials 22, no. 3: 1472-1514.
[25] Wang, Xiaofei i inni. 2023. „AIML TIG Technical Report Draft". IEEE 802.11 May 2023 Interim Meeting. Dostęp 17 czerwca 2023. https://mentor.ieee.org/802.11/dcn/22/11-22-0987-07-aiml-aiml-tig-technical-report-draft.doc.
[26] Wang, Zhou, Xuming Fang, Li Yan, Rong He, Yan Long. 2021. „Intelligent hybrid automatic repeat request retransmission for multi band Wi Fi networks." IET Communications 15, no. 9: 1249-1258.
[27] Wi-Fi Alliance. 2023. „Wi-Fi® by the numbers: Technology momentum in 2023". Dostęp 17 czerwca 2023. https://www.wi-fi.org/beacon/the-beacon/wi-fi-by-the-numbers-technology-momentum-in-2023.
[28] Wojnar, Maksymilian, Wojciech Ciężobka. 2023. „Zastosowanie algorytmów wielorękich bandytów do wyboru szybkości transmisji w sieciach IEEE 802.11ax". Konferencja Radiokomunikacji i Teleinformatyki KRiT 2023, Kraków, 20-22.09.2023.
[29] Wydmański, Witold, Szymon Szott. „Contention window optimization in IEEE 802.11ax networks with deep reinforcement learning." 2021 IEEE Wireless Communications and Networking Conference (WCNC).
[30] Zhang, Chen, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. 2021. „A survey on federated learning." Knowledge-Based Systems 216: 106775.

Pamiętaj o prenumeracie 2023

www.przegladtelekomunikacyjny.pl ***** www.sigma-not.pl

DOI: 10.15199/59.2023.4.2

Przeżywalność elastycznych sieci optycznych ze zwielokrotnieniem przestrzennym

Survivability of elastic optical networks with space division mutliplexing

STRESZCZENIE: Sieci teleinformatyczne stały się nieodzownym elementem naszego codziennego życia. Ich rosnąca popularność oraz rola w życiu społeczeństwa sprawiają, iż muszą one sprostać rosnącym wymaganiom, między innymi w zakresie efektywnego przesyłania dużych ilości danych oraz wysokiej dostępności i nieprzerwanej pracy. Niniejszy artykuł przedstawia narzędzia pozwalające sprostać tym oczekiwaniom – technologie nowoczesnych optycznych sieci transportowych oraz mechanizmy zapewnienia ich przeżywalności. Analiza teoretyczna uzupełniona jest przeglądem literaturowym wybranych istniejących/zaimplementowanych mechanizmów.
SŁOWA KLUCZOWE: sieci optyczne, transportowe sieci teleinformatyczne, przeżywalność sieci, elastyczne sieci optyczne, sieci optyczne ze zwielokrotnieniem przestrzennym

ABSTRACT: Recently, telecommunication networks have become an indispensable part of our every-day life. However, their increasing popularity and significant role in the society bring a growth of the requirements they have to meet, including an efficient support of tremendous traffic volumes and necessity of a high availability and provision of an uninterrupted and continuous operation. This manuscript discusses approaches allowing to meet the growing requirements – technologies of modern transport optical networks and survivability mechanisms. The theoretical analysis is enriched with the literature review on selected existing proposals of implementations.
KEYWORDS: optical networks, transport telecommunication networks, network survivability, elastic optical networks, space division multiplexing

dr hab. inż. RÓŻA GOŚCIEŃ
Katedra Systemów i Sieci Komputerowych,
Wydział Informatyki i Telekomunikacji,
Politechnika Wrocławska, Wrocław, Polska,
Roza.Goscien@pwr.edu.pl

Początki sieci teleinformatycznych datujemy na lata 60. XX wieku, kiedy to organizacja badawcza arpa stworzyła sieć arpanet. Była to prosta sieć oparta na transmisji pakietowej, która miała przede wszystkim wspierać wymianę plików i dzięki temu -- ułatwiać pracę badawczą organizacji. Sieć ta jednak bardzo szybko zaczęła się dynamicznie rozwijać, dając początki globalnej sieci połączeń zwanej dzisiaj Internetem [12]. Aktualnie sieci teleinformatyczne są nieodzownym elementem codziennego życia społeczeństwa, wspierając tak ważne i różnorodne dziedziny naszej aktywności jak praca, nauka, finanse, służba zdrowia, rozrywka czy też życie społeczne. Niezwykle ważną rolę sieci w naszym społeczeństwie podkreśliła pandemia covid-19, kiedy to wiele obszarów naszej działalności mogło funkcjonować jedynie w formie zdalnie – z pomocą sieci teleinformatycznych [14].

Ogromna popularność sieci i ich wszechobecność są również widoczne przy analizie trendów i statystyk związanych z rynkiem sieciowym. Firma Cisco przewiduje, iż liczba użytkowników internetowych wzrośnie od 3,9 miliarda w roku 2018 (51% światowej populacji) do 5,3 miliarda w roku 2023 (66% światowej populacji). Cisco szacuje również, iż średnia liczba urządzeń sieciowych przypadająca na jednego użytkownika sieci wzrośnie od 2,4 w roku 2018 do 3,6 do roku 2023. Co więcej, szacuje że średnia szybkość dostępowa łącza internetowego wzrośnie z 45,9 Mbps w roku 2018 do aż 110,4 Mbps w roku 2023 [1]. Wynikiem przedstawionych trendów, będzie niewątpliwie ogromny wzrost ruchu sieciowego. Przedstawione trendy wzrostowe stanowią szczególne wyzwanie w przypadku sieci transportowych, które przesyłają ruch zagregowany (od wielu użytkowników oraz urządzeń z konkretnych lokalizacji fizycznych) [1].

Należy zauważyć, iż rosnąca popularność sieci oraz ich wsparcie w kluczowych dla nas dziedzinach wiąże się również z oczekiwaniami zapewnienia ciągłej dostępności sieci oraz świadczenia ich usług w sposób bezpieczny dla użytkownika końcowego i jego danych. Sieci składają się jednak z urządzeń, które zawsze są podatne na awarie. Dodatkowo, sieci (albo ich węzły) mogą stać się celem różnych ataków, które mogą skutkować niedostępnością części fizycznej infrastruktury. Aby zminimalizować lub całkowicie wykluczyć niedostępność sieci powstającą w wyniku awarii lub ataku, niezbędne jest implementowanie w niej mechanizmów przeżywalnościowych [2].

Aby sprostać nowym oraz rosnącym wymaganiom, sieci teleinformatyczne muszą się stale rozwijać – zarówno w kwestii fizycznych architektur jak i programowych rozwiązań dedykowanych do sterowania ich operacjami. W przypadku sieci transportowych, fizyczne architektury opierają się na technologii optycznej, gdyż tylko światłowód jest w stanie obsłużyć transmisję tak dużych wolumenów danych na duże odległości (miedzy różnymi miastami, państwami a nawet kontynentami) z zapewnieniem wysokiej odporności na próby podsłuchu danych oraz zakłócenia ze środowiska zewnętrznego. Prace badawcze nad tymi architekturami przyniosły tak (ang. elastic optical networks [4], eon), zwielokrotnienie w dziedzinie przestrzennej (ang. *space division multiplexing*, sdm) [10] czy też architektura implementująca oba te rozwiązania jednocześnie – eon-sdm. Rozwój dedykowanych rozwiązań programowych do sterowania sieciami przynosi za to pomysły nowych algorytmów oraz całych scentralizowanych systemów do rozwiązywania problemów trasowania, przydziału zasobów sieciowych czy też zapewnienia ich przeżywalności [17].

W niniejszym artykule, przedstawione zostają najpopularniejsze dziś architektury optycznych sieci transportowych, wraz z omówieniem ich najważniejszych cech, zalet oraz wad. Następnie, omówione zostają mechanizmy przeżywalościowe możliwe do zaimplementowania w tych technologiach oraz przedstawione i przeanalizowane zostają przykładowe ich implementacje.

ELASTYCZNE SIECI OPTYCZNE ZE ZWIELOKROTNIENIEM PRZESTRZENNYM

Elastyczne sieci optyczne ze zwielokrotnieniem przestrzennym to architektura, która implementuje równocześnie dwa kluczowe rozwiązania – technologię eon i sdm. Niniejszy rozdział przedstawia tę architekturę, z uwzględnieniem jej najważniejszych cech, zalet oraz wyzwań.

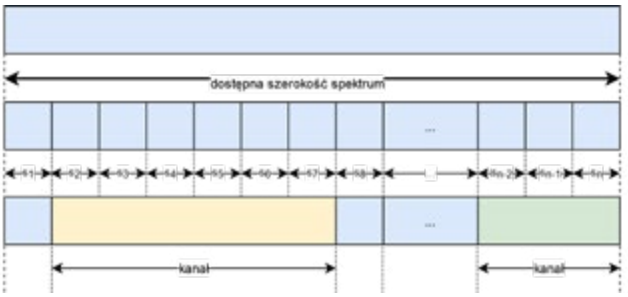
Elastyczne sieci optyczne

Architektura elastycznych sieci optycznych zaproponowana została w roku 2010 [11,15]. Aktualnie wdrożona jest częściowo lub całkowicie w wielu optycznych sieciach transportowych. Przez wielu uważana jest za następcę technologii wdm (ang. *wavelength division multiplexing*), którą znacznie przewyższa pod względem efektywności wykorzystania spektrum.

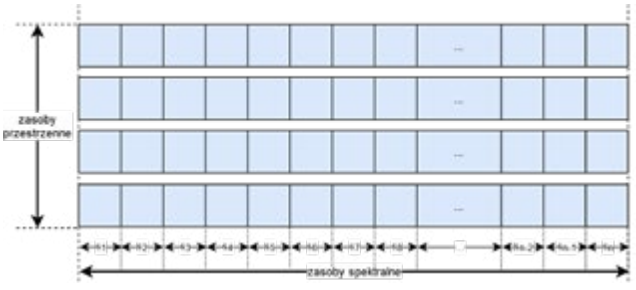
Podstawowa idea sieci eon polega na podziale spektrum na wąskie segmenty zwane slotami. Szerokość pojedynczego slotu to wg itu-t 6,25 lub 12,5 GHz. Sąsiednie sloty mogą być grupowane w kanały transmisyjne, dzięki czemu sieci te pozwalają dopasować szerokość kanału do wielkości obsługiwanych żądań. Schemat podziału zasobów spektralnych i tworzenia kanałów został zaprezentowany na *Rysunku 1*. Dodatkowo, technologia eon wspiera zaawansowane formaty modulacji (dobierane adaptacyjnie do wielkości żądania i długości ścieżki) oraz techniki transmisji (jak np. *orthogonal frequency division multiplexing*, ofdm). Dzięki temu, uzyskuje ona zdecydowanie wyższą efektywność spektralną niż technologia wdm.

Zwielokrotnienie w dziedzinie przestrzennej

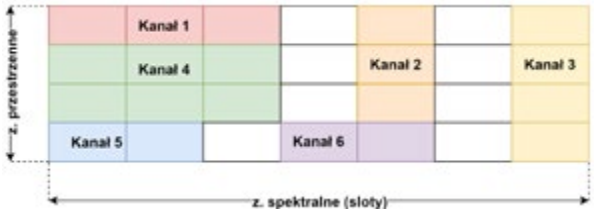
Architektura zwielokrotnienia w dziedzinie przestrzennej zakłada, iż każde fizyczne łącze to specjalnie zaprojektowany światłowód, składający się z wielu zasobów przestrzennych (ang. *spatial resources*) . Wspomniane zasoby przestrzenne mogą być fizycznie realizowane na różne sposoby (np. jako mody, rdzenie lub kable światłowodowe), a każdy z nich oferuje tę samą szerokość spektrum (ang. *spectral resources*). Spektrum w technologii sdm może zostać zorganizowane zgodnie ze sztywną siatką wdm lub elastyczną eon. W związku z tym, rozwiązanie sdm nie jest technologią niezależną, a współdziałającą z wdm lub eon [10,16]. *Rysunek* 2 przedstawia ideę sieci sdm przy podziale zasobów spektralnych wg reguł eon.



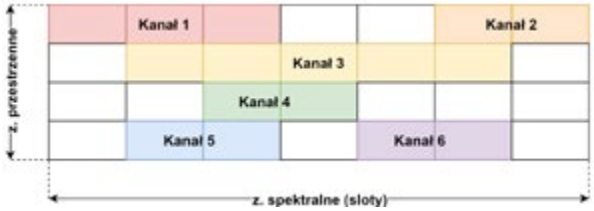
» *Rys. 1. Schemat podziału zasobów spektralnych oraz tworzenia kanałów w sieciach EON.*
» *Fig. 1. Spectrum division and channel's creation in EONS*



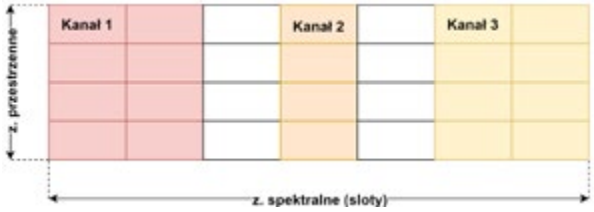
» *Rys. 2. Schemat podziału zasobów spektralnych w sieciach SDM/EON.*
» *Fig. 2. Spectrum division in EON/SDM*



(a) IndSw



(b) FracSw



(c) JSw

» *Rys. 3. Metody alokowania i przełączania kanałów w sieciach EON/SDM.*
» *Fig. 3. Swiching policies in EON/SDM*

W sieciach sdm dostępne są trzy sposoby przełączania zasobów przestrzennych, a co za tym idzie – tworzenia kanałów transmisyjnych. Pierwszą techniką jest przełączanie niezależne (ang. *independent switching*, IndSw), w którym możliwe jest tworzenie jedynie kanałów spektralnych (czyli ciągłych przedziałów częstotliwości umiejscowionych na jednym wybranym zasobie przestrzennym). Rozwiązanie to charakteryzuje się najwyższą efektywnością spektralną i wymaga najbardziej skomplikowanych (i drogich) urządzeń przełączających. Drugą opcją jest przełączanie łączne (ang. *joint switching*, JSw), w którym tworzone są kanały przestrzenno-spektralne (ciągłe przedziały częstotliwości zaalokowane na wszystkich zasobach przestrzennych równocześnie). Technika JSw zapewnia najniższą efektywność spektralną, ale równocześnie wymaga najprostszych (i najtańszych) urządzeń przełączających. Jako kompromis pomiędzy IndSw i JSw, powstała opcja przełączania grupowego (ang. *fraction swicthing*, FracSw), która pracuje z kanałami przestrzenno-spektralnymi, jednak zaalokowanymi tylko na podzbiorze zasobów przestrzennych. Dzięki temu, zarówno pod względem wykorzystania spektrum jak i wysmagań co do urządzeń, plasuje się pomiędzy IndSw i JSw [10]. Techniki przełączania i tworzenia kanałów w sieciach eon-sdm zostały zaprezentowane na *Rysunku 3.*

Problem trasowania oraz przydziału zasobów przestrzennych i spektralnych w sieciach EON-SDM

Podstawowym problemem optymalizacyjnym w sieciach eon-sdm jest problem trasowania oraz przydziału zasobów przestrzennych i spektralnych (ang. *routing, space and spectrum allocation*, rssa) [10]. Problem ten może być rozważany w postaci statycznej (inaczej problem projektowania/wymiarowania sieci) albo w postaci dynamicznej. Celem zadania projektowania sieci jest zaalokowanie danego zbioru żądań przy minimalizacji/maksymalizacji wartości funkcji celu. Kryterium optymalizacji jest tutaj najczęściej zdefiniowane jako wymagana szerokość spektrum w sieci. W przypadku problemu dynamicznego, rozważana jest działająca sieć, której zadaniem jest obsługiwanie pojawiających się żądań ruchu, tak aby w określonym oknie czasu zrealizować jak najwięcej żądań. Alokacja żądania w sieci eon-sdm wymaga wyznaczenia oraz zarezerwowania dla niego ścieżki optycznej. Ścieżka optyczna to połączenie ścieżki routingu (prowadzącej od węzła źródłowego żądania do docelowego) oraz kanału spektralnego lub przestrzenno-spektralnego (w zależności od techniki przełączania) dopasowanego do wielkości żądania oraz długości ścieżki.

PRZEŻYWALNOŚĆ SIECI TELEINFORMATYCZNYCH

Do ochrony sieci przed skutkami awarii oraz ataków, wykorzystać można mechanizmy przeżywalnościowe. W ogólnym przypadku dzielimy je na protekcję oraz restorację. Protekcja to działania podejmowane na etapie projektowania sieci. Ich idea polega na alokacji dodatkowych (nadmiarowych) zasobów, celem wykorzystania ich do realizacji transmisji w przypadku wystąpienia awarii określonego typu. Dla takiego schematu awarii, mechanizmy protekcji gwarantują odzyskanie 100% danych w bardzo krótkim czasie. Mechanizmy protekcji związane są jednak z wysokim wykorzystaniem zasobów sieciowych (do alokacji w warunkach normalnych oraz w przypadku awarii), co skutkuje wyższym kosztem implementacji oraz utrzymania sieci. Mechanizmy restoracji to natomiast algorytmy odzyskiwania łączności w sieci uruchamiane po wystąpieniu awarii/ataku. Ponieważ operują one na ograniczonych zasobach sieciowych, nie gwarantują odzyskania 100 % danych, a ich czas działania jest dłuższy niż mechanizmów protekcji. Są jednak rozwiązaniami tańszymi i niewymagającymi dodatkowych działań na etapie projektowania sieci (w porównaniu do protekcji).

Protekcja sieci

Protekcja w optycznych sieciach transportowych realizowana jest najczęściej za pomocą metod ochrony ścieżki. Metody te zakładają, iż dla każdego żądania w sieci wyznacza się więcej niż jedną ścieżkę optyczną. Jedna z tych ścieżek jest ścieżką podstawową (wykorzystywaną w normalnych warunkach), pozostałe – ścieżkami zapasowymi (ochronnymi). W zależności od ustawień systemu i wykorzystywanych technologii, ścieżki zapasowe mogą transmitować dane równocześnie ze ścieżką podstawową

lub zostać aktywowane dopiero po wystąpieniu awarii obejmującej ścieżkę podstawową. Protekcja ścieżek umożliwia zdefiniowanie wielu ścieżek zapasowych, jednak w rzeczywistych systemach najczęściej wykorzystywana jest konfiguracja z jedną ścieżką zapasową.

Metody protekcji projektowane są z myślą o ochronie przed konkretnym typem awarii w sieci – np. awarii dowolnego łącza, awarii pojedynczego węzła, awarii dwóch dowolnych łącz, itp. Adresowany schemat awarii determinuje liczbę ścieżek zapasowych oraz charakter relacji (tj. wymaganej rozłączności) pomiędzy ścieżkami podstawową i zapasowymi. Jeśli adresowany jest przypadek awarii dowolnego łącza w sieci, ścieżki te muszą być krawędziowo-rozłączne. Jeśli sieć ma być chroniona przed awarią dowolnych dwóch łącz, wymagane są dwie ścieżki zapasowe krawędziowo-rozłączne. Aby ochronić sieć przed awarią dowolnego węzła, ścieżka podstawowa i zapasową muszą być węzłowo-rozłączne.

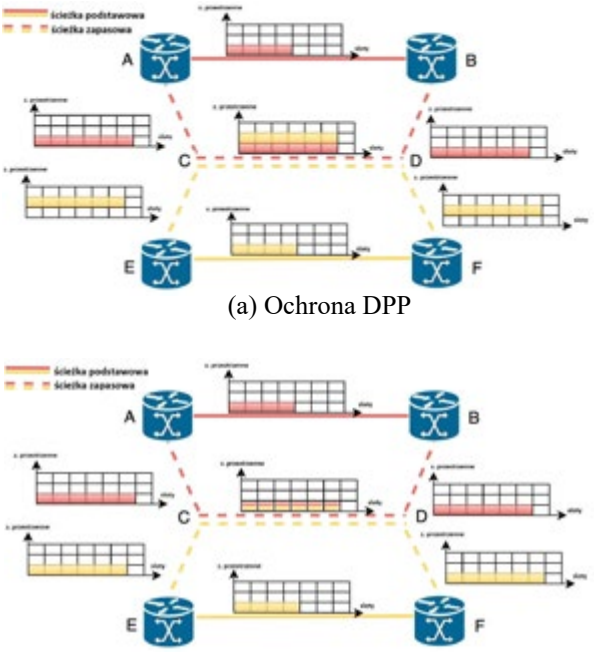
Pewną formą ochrony przepływów jest również routing wielościeżkowy, w którym żądanie rozkładane jest na wiele ścieżek realizowanych jednocześnie. Jeśli ścieżki wykazują odpowiedni stopień rozłączności, w przypadku awarii w sieci część żądania zawsze jest realizowana.

W sieciach optycznych protekcja ścieżek może działać w trybie dedykowanym lub współdzielonym. W przypadku dedykowanej protekcji (ang. *dedicated path protection*, dpp) zasoby optyczne (kanały częstotliwościowe) wyznaczone na każdej ścieżce (zarówno podstawowej jak i zapasowej) są dedykowane dla danego żądania i nie mogą być wykorzystywane przez żadne inne żądanie. We współdzielonej protekcji (ang. *shared backup path protection,* sbbp) zasoby ścieżek zapasowych mogą być współdzielone pomiędzy żądaniami, które nie ulegną awarii równocześnie. Na przykład, przy ochronie sieci przed awarią dowolnego łącza, zasoby ścieżek zapasowych mogę być współdzielone przez żądania, których ścieżki podstawowe są krawędziowo-rozłączne. Schematy dpp oraz sbpp dla sieci eon-sdm zostały zobrazowane na *Rysunku 4.*

Warto również wspomnieć, iż sieci eon wspierają częściową protekcję przepływów (ang. *squeezed protection,* sp), czyli schemat w którym chroniona jest tylko część przepływów. W takim przypadku, redukcji ulega rozmiar kanałów na ścieżkach zapasowych oraz całkowite wykorzystanie spektrum w sieci.

Restoracja danych

Mechanizmy restoracji w sieciach teleinformatycznych to próba zmian tras routingu dla żądań, których realizacja została przerwana przez awarię/atak. Proces ten podzielić można na kilka etapów. Pierwszym jest uaktualnienie informacji o stanie sieci, czyli zidentyfikowanie zasobów, które w wyniku awarii/ataku są niedostępne. Równocześnie identyfikowane są żądania ruchu, które z powodu niedostępności przydzielonych im zasobów nie mogą być realizowane i wymagają restoracji. Zwalniane są również zasoby dotychczas przydzielone takim żądaniom. Następnie, podejmowana jest próba realokacji tych żądań. W tym celu, może zostać zastosowany dowolny algorytm alokacji z uwzględnieniem uaktualnionej topologii sieci (z powodu awarii/ataku) oraz niedostępności zasobów przydzielonych żądaniom, których awaria/atak nie zakłóciły.



(a) Ochrona DPP

(b) Ochrona SBPP

» *Rys. 4. Schematy ochrony przepływów w sieciach eON-SDM–DPP oraz SBPP na podstawie [6].*
» *Fig. 4. Flow protection mechanisms in EON/SDM - DPP and SBPP based on [6].*

BADANIA EFEKTYWNOŚCI MECHANIZMÓW PRZEŻYWALNOŚCIOWYCH W SIECIACH EON-SDM

Większość istniejących prac na temat przeżywalnoości sieci eon-sdm skupia się na wykorzystaniu metod protekcji ścieżki – dedykowane [5,6,22] oraz współdzielonej [6,13,18]. Autorzy prac [5,6,13] studiowali statyczny problem projektowania sieci, do którego modelowania wykorzystali technikę programowania liniowego całkowitoliczbowego. Prace [5,6] rozważały ten problem z wykorzystaniem funkcji celu zdefiniowanej jako maksymalne wykorzystanie spektrum, natomiast praca [13] wykorzystała kombinację ilości zrealizowanego ruchu oraz wykorzystania spektrum. Dla tak zdefiniowanego problemu, zaproponowano liczne algorytmy heurystyczne [6] oraz metaheurystyczne [21]. Przykładowo, praca [21] proponuje oraz porównuje dedykowane implementacje algorytmów genetycznego oraz tabu search dla problemu projektowania sieci przy dedykowanej ochronie przepływów. W przypadku wykorzystania przeżywalności w rozwiązywaniu dynamicznych wersji problemu, funkcja celu głównie jest definiowana jako współczynnik odrzucenia żądań (ang. bandwidth blocking probability, bbp). Do rozwiązywania przypadków testowych proponowane i wykorzystywane są praktycznie tylko algorytmy heurystyczne. Należy tutaj zaznaczyć, iż większość istniejących badań adresuje problem ochrony przed awarią pojedynczego łącza w sieci [3,8]. Jedynie praca [19] skupia się na ochronie przed awarią dwóch łącz jednocześnie. Problem odzyskiwania przepływów w sieciach eon-sdm również poruszany był

w niewielu pracach. Autorzy [20] studiują problem restoracji przepływów w sieciach eon-sdm biorąc pod uwagę wyzwania płynące z istniejących rozwiązań światłowodowych oraz możliwości technologicznej realizacji sieci zdefiniowanych programowo z perspektywy kontrolera. Praca [7] skupia się na porównaniu różnych technik modelowania problemu restoracji w sieciach eon-sdm i na weryfikacji jego efektywności jako funkcji stopnia przeprojektowania sieci (tj. dostępności nadmiarowych zasobów).

W niniejszym rozdziale, przeanalizowane zostaną szczegółowo wybrane implementacje mechanizmów przeżywalnościowych w sieciach eon-sdm.

W przypadku wszystkich omawianych badań symulacyjnych, wykorzystany został model fizyczny sieci przedstawiony w pracy [9]. Zakłada on, iż transponder używa 3 slotów (37,5 GHz) oraz może stosować jedną z czterech modulacji: bpsk, qpsk, 16-qam, 32-qam. *Tabela 1* przedstawia wspierane wartości ruchu oraz odległości transmisji dla każdego formatu modulacji. Kiedy długość ścieżki przekracza maksymalny dystans konkretnej modulacji, wykorzystywane są regeneratory sygnału. Aby wybrać modulację dla konkretnego żądania oraz ścieżki routingu, stosowana jest zasada dat (ang. *distance adaptive transmission*) [23]. Wybiera ona najefektywniejszą spektralnie modulację, która równocześnie minimalizuje liczbę potrzebnych regeneratorów.

» *Tab. 1. Wspierana wielkość żądania oraz odległość transmisji dla transpondera pracującego na szeroko- ści spektrum 37,5 GHz [9]*
» *Tabl. 1. Supported bitrate and transmission reach for a transponder operating at 37.5 GHz spectrum [9].*

	Wielkość [Gbps]	Odległość [km]
BPSK	50	6300
QPSK	100	3500
32-QAM	150	1200
16-QAM	200	600

Protekcja w sieciach EON-SDM

Protekcja w sieciach jest spotykana przede wszystkim w zadaniu statycznego projektowania sieci dla zadanego zbioru ruchu sieciowego. Niemniej, stosowana jest również przy trasowaniu dynamicznym, kiedy to pojawiające się w sieci nowe żądania ruchu muszą być obsłużone z uwzględnieniem zasad przeżywalności.

Statyczny problem projektowania sieci z uwzględnieniem przeżywalności

W przypadku statycznego problemu projektowania sieci, należy zwymiarować sieć i/lub określić zasady routingu dla zdefiniowanego zbioru żądań ruchu. Do matematycznego opisu tego zagadnienia można wykorzystać jedną z dwóch notacji modelowania przepływów w sieciach (ścieżkowo-połączeniową (ang. *link-path*) lub węzłowo-połączeniową (ang. *node-link*)) oraz jedną z dwóch notacji modelowania alokacji kanałów w sieciach optycznych (opartą na kanałach (ang. *channel-based*) oraz opartą na slotach (ang. *slice-based*)) [24]. W niniejszym przykładzie użyte zostaną notacje ścieżkowo-połączeniowa oraz oparta na kanałach. Notacje te zostały wykorzystane w pracy [6] do opisu problemu projektowania sieci eon-sdm przy ochronie dpp oraz sbpp w następujący sposób.

Sieć modelowana jest jako graf skierowany (V, E, K), gdzie V to zbiór węzłów, E to zbiór skierowanych łącz oraz K to zbiór zasobów przestrzennych dostępnych na każdym łączu fizycznym. Zasoby spektralne na każdym zasobie przestrzennym zostają podzielone na sloty częstotliwościowe $s \in S$. Bazując na dostępnych slotach i zasobach przestrzennych, tworzony jest zbiór kanałów $c \in C$. Dla przełączania IndSw są to kanały spektralne, z których każdy opisany jest przez indeks zasobu przestrzennego, indeks pierwszego slotu oraz szerokość. W przypadku JSw kanały są przestrzenno-spektralne i przechodzą przez wszystkie zasoby przestrzenne. Każdy z nich opisany jest poprzez indeks pierwszego slotu i szerokość na pojedynczym zasobie przestrzennym. Dla opcji FracSw rozważane są kanały przestrzenno-spektralne, jednak obejmujące tylko podzbiór zasobów przestrzennych. Każdy kanał opisany jest poprzez indeks pierwszego slotu, szerokość na pojedynczym zasobie przestrzennym oraz indeksy wykorzystywanych zasobów przestrzennych.

Dany jest zbiór żądań ruchu $d \in D$, z których każde opisane jest poprzez indeks węzła początkowego, indeks węzła docelowego oraz wartość (w Gbps). Sieć chroniona jest przed awarią pojedynczego łącza, stąd każde żądanie musi zostać zrealizowane dwoma krawędziowo-rozłącznymi ścieżkami. Niech $p = (b_1,b_2)$ oznacza parę ścieżek routingu, przy czym b_1 jest ścieżką podstawową i b_2 jest ścieżką zapasową. Dla każdego żądania $d \in D$, dostępny jest zbiór n kandydujących par krawędziowo-rozłącznym ścieżek routingu $p = (b_1,b_2) \in P_d$.

Niech ścieżka-optyczna $l = (b, c)$ będzie połączeniem ścieżki routingu b oraz kanału c (spektralnego lub przestrzenno-spektralnego). Następnie, niech $q = (l_1,l_2)$ będzie parą ścieżek optycznych, opartą na parze krawędziowo-rozłącznych ścieżek routingu $p = (b_1,b_2)$. Para ścieżek optycznych pozwala zaalokować żądanie z uwzględnieniem ochrony dpp lub sbpp. W niniejszym przykładzie, dla każdego żądania ruchu $d \in D$ wyliczany jest zbiór kandydujących par ścieżek optycznych $q = (l_1,l_2) \in Q_d$. Bazuje on na kandydujących parach ścieżek routingu $(p = (b_1,b_2) \in P_d)$ oraz dostępnym zbiorze kanałów $c \in C$. Warto zaznaczyć, iż dla każdego żądania rozważane są tylko kanały o rozmiarze dopasowanym do jego wymagań (zgodnie z regułą dat).

W pracy [6] przedstawione zostały generyczne modele opisu problemu projektowania sieci eon-sdm z funkcją celu zdefiniowaną jako wykorzystanie spektrum w sieci. Pozwalają one uwzględnić różne metody ochrony (dpp, sbpp, częściowa ochrona sieci) oraz różne techniki przełączania (IndSDw}, FracSw, JSw). Modele wykorzystują notację opartą na kanałach, która pozwala w łatwy sposób modelować różne techniki przełączania. Ich wybór determinuje definicję kanałów (spektralne vs. przestrzenno-spektralne) oraz liczbę wykorzystywanych przez nie zasobów przestrzennych. Ta sama notacja pozwala również łatwo modelować częściową

ochronę przepływów, poprzez alokowanie kanałów zapasowych dopasowanych do wielkości ruchu podlegającego ochronie.

Oznaczenia:

$e \in \mathrm{E}$	łącza w sieci
$s \in S$	sloty częstotliwościowe
$k \in \mathrm{K}$	zasoby przestrzenne
$d \in \mathrm{D}$	żądania ruchu
$c \in C$	kanały częstotliwościowe
$p = (b_1,b_2) \in P_\mathrm{d}$	pary kandydujących krawędziowo rozłącznych ścieżek routingu dla żądania d
$l = (b, c)$	ścieżka optyczna (połączenie kanału i ścieżki routingu)
$q = (l_1,l_2) \in \mathrm{Q_d}$	kandydujące pary ścieżek optycznych dla żądania d

Stałe:

α_{dqsk}	= 1, jeśli podstawowa ścieżka optyczna z pary ścieżek q dla żądania d używa slotu s na zasobie przestrzennym k; 0, w przeciwnym razie
α'_{dqsk}	= 1, jeżeli zapasowa ścieżka optyczna z pary ścieżek q dla żądania d używa slotu s ona zasobie przestrzennym k; 0, w przeciwnym razie
β_{edq}	= 1, jeżeli podstawowa ścieżka optyczna pary q dla żądania d używa łącza e; 0, w przeciwnym razie
β'_{edq}	= 1, jeżeli zapasowa ścieżka optyczna z pary q dla żądania d używa łącza e; 0, w przeciwnym razie

Zmienne (binarne):

x_{dq}	=1, jeżeli żądanie d jest realizowane z wykorzystaniem pary q; 0, w przeciwnym razie
y_{esk}	=1, jeżeli slot s jest używany na zasobie przestrzennym k na łączu e; 0, w przeciwnym razie
y_{es}	=1, jeżeli slot s jest używany na którymkolwiek zasobie przestrzennym na łączu e; 0, w przeciwnym razie
y_s	=1, jeżeli slot s jest używany na którymkolwiek łączu; 0, w przeciwnym razie
w_{esk}	=1, jeżeli slot s jest używany na zasobie przestrzennym k na łączu e w ramach podstawowej ścieżki optycznej; 0, w przeciwnym razie
z_{egsk}	=1, jeżeli slot s jest używany na zasobie przestrzennym k na łączu e w przypadku awarii łącza g; 0, w przeciwnym razie

A) Dedykowana protekcja ścieżek

Problem alokacji żądań w sieci eon-sdm przy uwzględnieniu dedykowanej protekcji ścieżek opisany jest funkcją celu (1) oraz ograniczeniami (2)–(5). Równanie (2) gwarantuje, że dla każdego żądania wybrana zostaje dokładnie jedna para ścieżek optycznych. Ograniczenie (3) zapewnia, że pojedynczy slot na konkretnym zasobie przestrzennym i łączu fizycznym zostaje przydzielony maksymalnie do jednego żądania. Ostatnie dwa równania, (4) oraz (5), kontrolują wykorzystanie spektrum oraz definiują zmienne y_{es} i y_s.

Funkcja celu:

$$\min y = \sum_s y_s . \tag{1}$$

Ograniczenia:

$$\sum_{q \in Q_d} x_{dq} = 1, d \in D. \tag{2}$$

$$\sum_{d \in D} \sum_{q \in Q_d} x_{dq}\left(\alpha_{dqsk}\beta_{edq} + \alpha'_{dqsk}\beta'_{edq}\right) \le y_{esk},$$

$$e \in E, s \in S, k \in K. \tag{3}$$

$$y_{esk} \le y_{es}, e \in E, s \in S, k \in K. \tag{4}$$

$$y_{es} \le y_s, e \in E, s \in S. \tag{5}$$

B) Współdzielona protekcja ścieżek

Problem ochrony przy rozważaniu współdzielonej protekcji ścieżek opisany jest poprzez funkcję celu (6) oraz ograniczenia (7)-(12). Równanie (7) zapewnia, że dla każdego żądania wyznaczona zostaje dokładnie jedna para ścieżek optycznych. Ograniczenia (8)-(10) kontrolują wykorzystanie spektrum oraz jego współdzielenie, które jest możliwe tylko pomiędzy żądaniami, których podstawowe ścieżki nie mają wspólnych łącz. Ostatnie równania, (11) i (12), odpowiedzialne są za kontrolowanie wykorzystania spektrum oraz definiowania zmiennych y_{es} i y_s.

Funkcja celu:

$$\min y = \sum_s y_s. \tag{6}$$

Ograniczenia:

$$\sum_{q \in Q_d} x_{dq} = 1, d \in D. \tag{7}$$

$$\sum_{d \in D} \sum_{q \in Q_d} x_{dq}\alpha_{dqsk}\beta_{edq} = w_{esk},$$

$$e \in E, s \in S, k \in K. \tag{8}$$

$$\sum_{d \in D} \sum_{q \in Q_d} x_{dq}\beta_{gdq}\alpha'_{dqsk}\beta'_{edq} = z_{egsk},$$

$$e, g \in E, e \ne g, s \in S, k \in K. \tag{9}$$

$$w_{esk} + z_{egsk} \leq y_{esk}, e \in E, s \in S, k \in K. \quad (10)$$

$$y_{esk} \leq y_{es}, e \in E, s \in S, k \in K. \quad (11)$$

$$y_{es} \leq y_s, e \in E, s \in S. \quad (12)$$

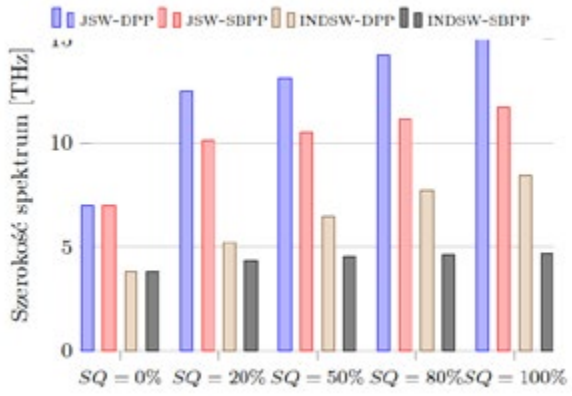
C) Efektywność statycznego projektowania sieci z uwzględnieniem przeżywalności

Dla rozwiązania rzeczywistych przypadków testowych oraz przeprowadzenia badań symulacyjnych weryfikujących efektywność przeżywalności, w pracy [6] zaproponowano również wysoce-skalowalny algorytm heurystyczny. Z jego pomocą przeprowadzono badania mające na celu, między innymi, zweryfikowanie jak dużo zasobów spektralnych jest niezbędnych do zrealizowania ochrony w sieci eon-sdm, w zależności od stosowanej metody ochrony (dpp, sbpp), ilości chronionych przepływów (współczynnik **sq**, od 0% (brak ochrony) do 100% (ochrona wszystkich danych) z krokiem 10%) oraz metody przełączania (IndSw, JSw). Przykładowe wyniki zaprezentowane są na *Rysunku 5*. Wyniki pokazują, iż wykorzystanie spektrum w sieci zależy przede wszystkim od ilości chronionych przepływów – im większa wartość współczynnika sq, tym większe zużycie spektrum. Jest to szczególnie widoczne dla ochrony dedykowanej, gdzie do całkowitej ochrony przepływów (**sq**=100%) niezbędne jest zaalokowanie ponad dwa razy tyle spektrum co do alokacji przepływów bez żadnej ochrony (**sq**=0%). W przypadku ochrony współdzielonej wymagana ilość spektrum jest dużo mniejsza. Całkowita ochrona przepływów wymaga ok. 23% więcej spektrum dla IndSw oraz 67-69% dla JSw, w porównaniu do sieci niechronionej. Wyniki te pokazują równocześnie, iż technika przełączania IndSw jest dużo bardziej efektywna spektralnie, niż JSw.

Dynamiczna alokacja żądań z uwzględnieniem przeżywalności

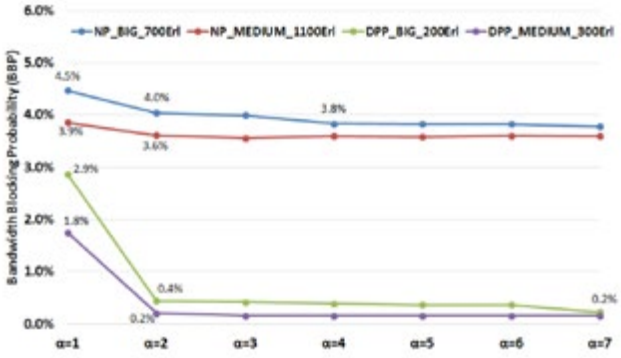
Dynamiczna alokacja to przydzielanie ścieżek optycznych na bieżąco dla pojawiających się w sieci żądań oraz zwalnianie ich po zrealizowaniu żądania. Ze względu na ograniczoną ilość zasobów sieciowych oraz pojawiającą się z czasem fragmentację łącza, sieć może nie być w stanie obsłużyć wszystkich żądań. Celem problemu jest więc maksymalizacja sumy zrealizowanych żądań, bądź mówiąc inaczej – minimalizacja sumy żądań odrzuconych. Formalnie, problemy dynamiczne rozważa się z funkcją celu zdefiniowaną jako prawdopodobieństwo blokady żądania (ang. *bandwidth blocking probability*, bbp), czyli stosunek sumy odrzuconych żądań do wszystkich zaoferowanych.

Problem dynamicznej alokacji żądań w sieci eon-sdm z uwzględnieniem przeżywalności został zaadresowany w pracach [3,8]. Do realizacji przeżywalności, Autorzy wykorzystali dpp w konfiguracji *jedna ścieżka podstawowa – jedna ścieżka zapasowa*. Do rozwiązania problemu zaproponowali kilka algorytmów heurystycznych i porównali ich efektywność. Następnie, wykorzystali najlepszy z nich do wykazania, iż wykorzystanie kanałów przestrzenno-spektralnych może podnieść efektywność routingu dynamicznego w przeżywalnej sieci eon-sdm. Spowodowane jest to faktem, iż



» *Rys. 5. Wymagana szerokość spektrum w sieci Euro28 dla sumarycznego ruchu 600 Tbps oraz |K|=7 zasobów spektralnych [6].*
» *Fig 5. Spectrum width in Euro28, total traffic volume 600 Tbps and |K|=7 spatial resources [6].*

w przypadku obciążonej sieci (realizacją żądań) oraz wysokiej fragmentacji spektrum może być łatwiej znaleźć kanał przestrzenno-spektralny o odpowiedniej pojemności niż kanał spektralny o tej samej pojemności. Pojemność kanału przestrzenno-spektralnego to jego szerokość na pojedynczym zasobie przestrzennym pomnożona przez liczbę wykorzystywanych zasobów przestrzennych, a pojemność kanału spektralnego to bezpośrednio jego szerokość na wybranym zasobie przestrzennym. *Rysunek 6* przedstawia bbp w funkcji parametru α , czyli dopuszczalnego rozmiaru przestrzennego kanałów. Analiza została przeprowadzona dla sieci niechronionej (ang. *no protection,* np) oraz implementującej ochronę dpp. Rozważono również różne wartości natężenia ruchu w sieci (200 Erl, 300 Erl, 700 Erl, 1100 Erl) oraz dwa scenariusze wielkości pojedynczych żądań ruchu (medium – średnia wartość żądania 500 Gbps, big – średnia wartość żądania 1000 Gbps). Wyniki wyraźnie pokazują korzyści z zastosowania kanałów przestrzenno-spektralnych, które pozwalają zauważalnie zredukować bbp w każdej z rozważanych konfiguracji. Zanotowane korzyści były szczególnie wysokie dla sieci implementującej dpp, gdzie dopuszczenie kanałów o szerokości spektralnej 2 (w stosunku do samych kanałów spektralnych) pozwoliło zredukować bbp o nawet 2,5 punktu proc. (dla dużych żądań).



» *Rys. 6. Prawdopodobieństwo blokady żądania w sieci eon-sdm jako funkcja rozmiaru przestrzennego (α) dostępnych kanałów przy |K|=7 zasobach przestrzennych [3].*
» *Fig 6. Bandwidth blocking probability in eon-sdm network with |K|=7 spatial resources as a function of channels' space dimension (α) [3].*

Restoracja w sieciach EON-SDM

Problem restoracji rozważany jest w sieci działającej, w której awaria lub atak zakłóciły realizację części żądań. Żądanie niezakłócone cały czas pozostają realizowane w sieci, a te zakłócone – podlegają restoracji (próbie odzyskania/realokacji). Pierwszym krokiem restoracji jest uaktualnienie informacji o sieci (o zasobach niedostępnych z powodu awarii/ataku lub rezerwacji dla żądań niezakłóconych), identyfikacji żądań podlegających realokacji oraz zwolnienie przydzielonych ich wcześniej zasobów. Następnie, podejmowana jest próba restoracji tych żądań w uaktualnionej topologii sieci.

Modelowanie problemu restoracji

Model restoracji przepływów po awarii w sieci eon-sdm zaproponowany został w pracy [7]. Opisany jest on za pomocą funkcji celu (13) oraz ograniczeń (14), (15). Celem problemu jest maksymalizacja sumarycznej wartości odzyskanych żądań.

Oznaczenia:

$e \in E'$	łącza sieci dostępne w zaktualizowanej topologii (po awarii)
$s \in S'$	sloty częstotliwościowe dostępne na każdym zasobie przestrzennym podczas procesu restoracji
$d \in D'$	żądania ruchu wymagające restoracji
$c \in C'$	kanały częstotliwościowe dla zbioru S' slotów
$p \in P'_d$	kandydujące ścieżki routing dla żądania d w uaktualnionej topologii sieci
$l = (b, c) \in L'_d$	kandydujące ścieżki optyczne dla żądania d w uaktualnionej topologii sieci

Stałe:

b_d	wielkość (w Gbps) żądania d
ρ	współczynnik przewymiarowania – określa ile razy więcej slotów jest dostępnych w sieci niż wymagane minimum do obsługi żądań w normalnym stanie
δ_{eks}	dla konkretnego stanu sieci: =1, jeśli slot s jest zajęty na zasobie przestrzennym k na łączu e; 0, w przeciwnym razie
γ_{le}	=1, jeśli ścieżka optyczna l wykorzystuje łącze e; 0, w przeciwnym razie
α_{cks}	=1, jeśli kanał c używa slotu s na zasobie przestrzennym k; 0, w przeciwnym razie

Zmienne (binarne):

z_{dl}	=1, żądanie d jest realizowane za pomocą ścieżki optycznej l; 0, w przeciwnym razie
r_d	=1, jeśli żądanie d zostaje odrzucone z powodu awarii/ataku; 0, w przeciwnym razie
y_{eks}	=1, jeśli slot s jest używany na zasobie przestrzennym k na łączu e; 0, w przeciwnym razie

Funkcja celu:

$$\max \sum_{d \in D'} (1 - r_d) \cdot b_d \tag{13}$$

Ograniczenia:

$$\sum_{l \in L'_d} z_{dl} + r_d = 1, d \in D'. \tag{14}$$

$$\sum_{d \in D'} \sum_{l \in L'_d} z_{dl} \cdot \gamma_{le} \cdot \alpha_{c(l)ks} + \delta_{eks} = y_{eks}, \tag{15}$$

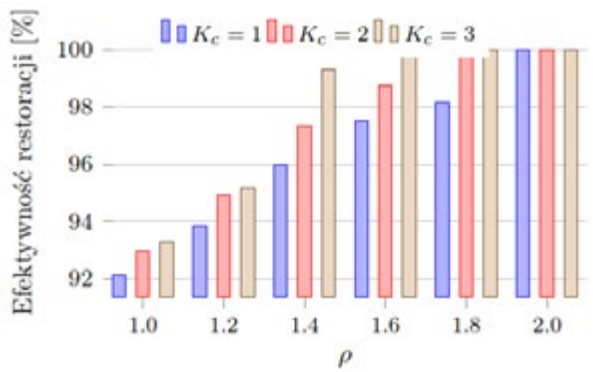
$$e \in E', k \in K, s \in S'.$$

Ograniczenie (14) wskazuje, iż każde żądanie zostaje albo zrealizowane (ma wyznaczoną działającą ścieżkę optyczną) albo odrzucone. Następnie, równanie (15) kontroluje wykorzystanie spektrum oraz jego nienakładanie się. Formuła ta bierze pod uwagę żądania zaalokowane wcześniej w sieci oraz te, które podlegają restoracji.

Efektywność procesu restoracji

Podstawowym kryterium oceny mechanizmów restoracji jest ilość danych, które udało się odzyskać. Jeśli sieć jest zwymiarowana idealnie pod obsługiwany ruch, proces restoracji może wykazać się niską skutecznością. Stąd też, sieci często projektowane są z nadmiarem, tak aby zapewnić dodatkowe zasoby na obsługę rosnącego ruchu lub na wypadek restoracji.

Badania nad efektywnością procesu restoracji w sieciach eon-sdm prowadzone były w pracy [7]. Jednym z celów pracy było zbadanie efektywności tego procesu (mierzonej jako procent odzyskanych danych) w funkcji wielkości przewymiarowania sieci ρ. Innymi słowy, parametr określa w tym badaniu ile razy więcej zasobów spektralnych (slotów częstotliwościowych) posiada sieć w stosunku do tego, co jest wymagane do obsługi ruchu w normalnych warunkach. Wyniki przedstawione zostały dla sieci eon-sdm zakładając liczbę zasobów przestrzennych równą $|K|=4$ oraz możliwość tworzenia kanałów przestrzennych grupujących $K_c=1,2,3$ zasoby przestrzenne. Przykładowe wyniki zaprezentowane są na *Rysunku 7*. Pokazują one, iż w przypadku sieci



» *Rys. 7. Efektywność restoracji jako funkcja parametru ρ dla topologii PL12, całościowego ruchu sieciowego 15 Tbps [7].*
» *Fig 7. Restoration efficiency as a function of ρ for pl12 network with overall traffic 15 Tbps [7].*

zwymiarowanej idealnie pod obsługiwany ruchu efektywność restoracji wynosi ok. 92-93%. Pełne odzyskanie danych wymaga dodatkowych zasobów. W badanym przypadku, niezbędne było nawet i dwa razy tyle zasobów, aby odzyskać 100% danych. Warto również zauważyć, iż proces restoracji jest znacznie trudniejszy dla sieci ze zwielokrotnieniem przestrzennym ($|K|>1$), gdy możliwe jest tworzenie kanałów przestrzenno-spektralnych. Przyczyną jest problem znalezienia takiego kanału w sieci, która już obsługuje pewien zbiór żądań.

PODSUMOWANIE

Sieci teleinformatyczne są aktualnie nieodzownym elementem życia społeczeństwa. Ta istotna rola pociąga za sobą jednak potrzebę ich ciągłego rozwoju, celem obsługi rosnącego ruchu sieciowego oraz zapewnienia ciągłości i bezpieczeństwa ich pracy. Wysoka wydajność sieci oraz obsługa rosnącego ruchu mogą być zaadresowane przez nowe architektury i technologie sieciowe, spośród których aktualnie najwięcej uwagi naukowców i operatorów przyciąga idea elastycznych sieci optycznych ze zwielokrotnieniem przestrzennym. Ciągła praca sieci i ich niezawodność wymagają natomiast implementacji mechanizmów przeżywalnościowych. Mechanizmy te dzielimy na protekcję (dedykowaną oraz współdzieloną) i restorację. Niniejszy artykuł przestawił i przedyskutował wybrane implementacje mechanizmów przeżywalnościowych w elastycznych sieciach optycznych ze zwielokrotnieniem przestrzennym. Wyniki przedstawionych badań jednoznacznie pokazują, iż mechanizmy przeżywalnościowe wpływają na efektywność sieci optycznej, przede wszystkim pod względem wykorzystania spektrum. Stąd też, powinny być one indywidualnie dopasowane do wymaganego stopnia ochrony oraz posiadanych zasobów. ●

LITERATURA

[1] Cisco. Cisco annual internet report 2018–2023, March 2020.

[2] T. Gomes et al. A survey of strategies for communication networks to protect against large-scale natural disasters. In *International Workshop on Resilient Networks Design and Modeling* (RNDM), 2016.

[3] R. Goścień. On the efficient dynamic routing in spectrally-spatially flexible optical networks. In *Proceedings of Resilient Networks Design and Modeling* (RNDM), 2019.

[4] R. Goścień et al. Protection in elastic optical networks. *IEEE Network,* 29(6):88–96, 2015.

[5] R. Goścień and P. Lechowicz. Column generation technique for optimization of survivable flex-grid SDM networks. In *Proc. of Resilient Networks Design and Modeling* (RNDM), 2017.

[6] R. Goścień and K. Walkowiak. On the efficiency of survivable flex-grid SDM networks. *Journal of Lightwave Technology,* 36(10):1815–1823, 2018.

[7] Róża Goścień. On the efficient flow restoration in spectrally-spatially flexible optical networks. *Electronics*, 10(12), 2021.

[8] Róża Goścień and Paweł Ksieniewicz. Efficient dynamic routing in spectrally-spatially flexible optical networks based on traffic categorization and super- vised learning methods. *Optical Switching and Ne tworking,* 43:100650, 2022.

[9] P. S. Khodashenas, D. Siracusa J. M. Rivas-Moscoso, F. Pederzolli, B. Shariati, D. Klonidis, E. Salvadori, and I. Tomkos. Comparison of spectral and spatial super-channel allocation schemes for SDM networks. *Journal of Lightwave Technology,* 34(11):2710–2716, 2016.

[10] M. Klinkowski, P. Lechowicz, and K. Walkowiak. Survey of resource allocation schemes and algorithms in spectrally-spatially flexible optical networking. *Optical Switching and Networking*, 27:58–78, 2018.

[11] D. Klonidis, F. Cugini, O. Gerstel, M. Jinno, V. Lopez, E. Palkopoulou, M. Sekiya, D. Siracusa, G. Tho- uénon, and C. Betoule. Spectrally and spatially fle- xible optical network planning and operations. *IEEE Communications Magazine,* 53(2):69–78, 2015.

[12] Alberto Leon-Garcia and Inra Widjaja. *Communication Networks.* McGraw-Hill Education; 2nd edition, 2003.

[13] Y. Li, N. Hua, and X. Zheng. Routing, wavelength and core allocation planning for multi-core fiber networks with MIMO-based crosstalk suppression. In *Proc. of Opto-Electronics and Communications Conference* (OECC), 2015.

[14] Andra Lutu et al. A characterization of the covid-19 pandemic impact on a mobile network operator traffic. In *ACM Internet Measurement Conference,* 2020.

[15] M.Jinno, B. Kozicki, H. Takara, A. Watanabe, Y. So- ne, T. Tanaka, and A. Hirano. Distance-adaptive spectrum resource allocation in spectrum-sliced elastic optical path network. IEEE Communications Magazine, 48(8):138–145, 2010.

[16] A. Muhammad, G. Zervas, and R. Forchheimer. Resource allocation for space-division multiplexing: optical white box versus black box networking. *Journal of Lightwave Technology,* 33(23):4928–4941, 2015.

[17] Francesco Musumeci et al. An overview on application of machine learning techniques in optical networks. *IEEE Communications Surveys Tutorials*, 21(2):1383–1408, 2019.

[18] H. M. N. S. Oliveira and N. L. S. da Fonseca. Algorithm for shared path for protection of space division multiplexing elastic optical networks. In *Proc. of IEEE International Conference on Communications* (IEEE ICC), 2017.

[19] H. M. N. S. Oliveira and N. L. S. da Fonseca. Sha- ring spectrum and straddling p-cycle FIPP for protection against two simultaneous failures in SDM elastic optical networks. In *Proc. of IEEE Latin- American Conference on Communications* (IEEE LATINCOM), 2017.

[20] F. Pederzolli, M. Gerola, A. Zanardi, X. Forns, J. F. Ferran, and D. Siracusa. YAMATO: The first SDN control plane for independent, joint, and fractional - joint switched SDM optical networks. *Journal of Lightwave Technology*, 35(8):1335–1341, 2017.

[21] Michal Witold Przewozniczek, Róża Goścień, Piotr Lechowicz, and Krzysztof Walkowiak. Metaheuristic algorithms with solution encoding mixing for effective optimization of sdm optical networks. *Engineering Applications of Artificial Intelligence*, 95:103843, 2020.

[22] Y. Tan, R. Zhu, H. Yang, Y. Zhao, J. Zhang, Z. Liu, Q. Qu, and Z. Zhou. Crosstalk-aware provisioning strategy with dedicated path protection for elastic multi-core fiber networks. In *Proc. of International Conference on Optical Communications and Networks* (ICOCN), 2016.

[23] K. Walkowiak. *Modeling and optimization of cloud- ready and content--oriented networks.* Springer, 2016.

[24] K. Walkowiak, P. Lechowicz, M. Klinkowski, and A. Sen. ILP modeling of flexgrid SDM optical networks. In *Proc. of International Telecommuni- cations Network Strategy and Planning Symposium* (Networks), 2016.

DOI: 10.15199/59.2023.4.3

Ultra-szybka transmisja światłowodowa na potrzeby centrów danych

Ultra-highs speed optical fibre transmission in data centers

dr hab. inż. JAROSŁAW TURKIEWICZ, prof. uczelni
Instytut Telekomunikacji, Politechnika Warszawska,
Warszawa,
jaroslaw.turkiewicz@pw.edu.pl

STRESZCZENIE: Wzrost ilości informacji przesyłanych w ramach centrów danych i superkomputerów stymuluje rozwój i wykorzystanie dopasowanych do ich potrzeb oraz wymagań technik transmisyjnych. W referacie zostaną zaprezentowane techniki transmisji danych o wysokich przepływnościach wykorzystujące lasery VCSEL i światłowody wielomodowe.
SŁOWA KLUCZOWE: datainterconnect, laser, światłowód wielomodowy, transmisja, VCSEL

ABSTRACT: Growth of data transmitted within the data centers and supercomputers stimulates development of novel data transmission techniques adjusted to the data center needs. In the paper, ultra-high speed transmission techniques utilizing VCSEL laser and multimode fibres are presented.
KEYWORDS: datainterconnect, laser, multimode fibre, transmission, VCSEL.

Wzrost ruchu w sieciach teleinformatycznych, centrach danych jak i rozwój takich technologii jak 6G wymusza rozwój infrastruktury sprzętowej zdolnej do przesyłania ciągle rosnącej ilości danych. Kluczowym medium transmisyjnym do zaspokojenia zapotrzebowania na pasmo transmisyjne jest światłowód. O znaczeniu światłowodu dla rozwoju sieci Internet, gospodarki jak i współczesnej cywilizacji, świadczy przyznanie w roku 2006 Charlesowi Kao, Nagrody Nobla w dziedzienie fizyki *"for groundbreaking achievements concerning the transmission of light in fibers for optical communication"* [1]. W przeszłości liczba dostępnych światłowodowych technik transmisyjnych była ograniczona, dominowała transmisja z zwielokrotnieniem w dziedzinie długości fal WDM (*Wavelength Division Multiplexing*) i przepływności do 10 Gb/s. Obecnie obserwujemy trend rozwoju i dopasowywania technik transmisji do obszaru zastosowań i jego wymagań.

Komercyjnie dostępne systemy długodystansowe wykorzystują zaawansowane formaty modulacji, np. 16-QAM, gęste zwielokrotnienia w dziedzinie długości fal DWDM (*Dense Wavelength Division Multiplexing*) zwielokrotnienie polaryzacyjne, transmisję koherentną i zaawansowane cyfrowe przetwarzania sygnału w celu korekcji zniekształceń sygnału. Pozwala to na uzyskanie przepływności w pojedynczym kanale transmisyjnym do 1,6 Tbit/s [2], co w połączeniu z techniką DWDM umożliwia transmisję rzędu kilkudziesięciu Tbit/s w pojedynczym włóknie światłowodowym. Ze względu na minimum tłumienia

światłowodu jednomodowego (0,2 dB/km) systemy transmisji długodystansowej wykorzystują okno 1550 nm. Systemy dostępowe FTTH (*Fibre to the Home*) wykorzystują binarną modulację amplitudy o przepływności do 10 Gb/s. Wykorzystanie dość prostej techniki transmisji światłowodowej w systemach FTTH, przekłada się na niski koszt systemu, który jest dzielony na ograniczoną liczbę użytkowników, np. 32. Systemy te wykorzystują okna 1550 nm i 1310 nm.

Osobną klasę systemów transmisyjnych tworzą systemy wykorzystywane do transmisji w centrach danych jak i superkomputerach, nazywane datainterconnect. Systemy te wykorzystywane są do transmisji danych pomiędzy urządzeniami w ramach budynku czy stojaka i służą do transmisji danych na bardzo krótkie odległości do 2 km i przepływnościach rzędu kilkudziesięciu Gb/s. Jednym z wymagań na te systemy jest ograniczone opóźnienie (*latency*), wymagane do minimalizacji czasu przesyłania i przetwarzania informacji. Kluczowe są też niskie koszty kapitałowe i operacyjne ze względu na bardzo dużą liczbą połączeń w ramach jednego centrów danych, mogącą sięgać kilkudziesięciu tysięcy połączeń [3]. Redukcja zużycia energii i powiązana z tym minimalizacja kosztu operacyjnego możliwa jest dzięki wykorzystaniu technik transmisji w oknie 850 nm.

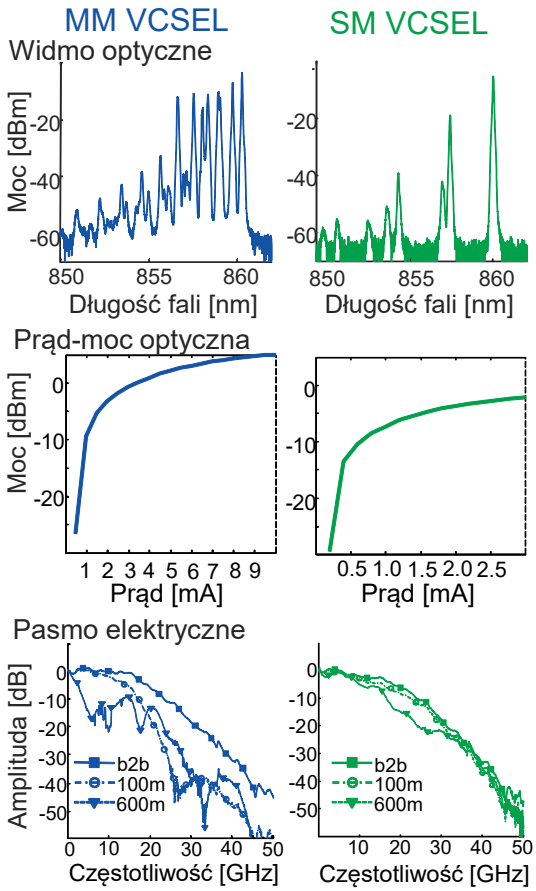
BUDOWA DATAINTERCONNECT

Rys. 1 pokazuje schemat budowy systemu transmisyjnego o wysokich przepływnościach typu datainterconnect. W tychże systemach wykorzystuje się jako medium transmisyjne światłowody wielomodowe MMF (*Multi Mode Fibre*), a jako źródło sygnału wykorzystuje się lasery z pionową wnęką rezonansową VCSEL (*Vertical Cavity Emitting Laser*). Światłowody MMF typu OM4 w oknie 850 nm charakteryzują się tłumieniem do 3 dB/km i pasmem modowym 4700 MHz·km [4]. Biorąc pod uwagę, długość połączeń w ramach centrów danych, gdzie ich większość jest poniżej 3m [3], tłumienie światłowodu nie ogranicza zasięgu transmisji dla bardzo krótkich połączeń.

» *Rys. 1. Generyczny schemat budowy datainterconnect*
» *Fig. 1 Generic datainterconnect scheme*

Pasmo modowe światłowodu OM4 umożliwia transmisję sygnałów o przepływnościach do 100 Gb/s na dystansie do 100 m. Czynnikiem ograniczającym zasięg transmisji dla wysokich przepływności może być dyspersja chromatyczna, która sięga -100 ps/nm·km w oknie 850 nm [5]. Najnowsza generacja światłowodów wielomodowych OM5 [6] ma parametry transmisyjne na długości fali 850 nm praktycznie takie same jak OM4. Cechą wyróżniającą OM5 jest możliwość wykorzystania zakresu fal 846 nm – 953 nm do transmisji wielokanałowej o wysokich przepływnościach SWDM (*Short Wavelength Division Multiplexing*). Światłowody OM4 nie gwarantują tego. Średnica rdzenia światłowodu OM4 i OM5 jest równa 50 µm, co ułatwia łączenie ze źródłem i odbiornikiem światła, a przez to znacząco obniża koszt montażu urządzenia.

W datainteconnect jako źródło sygnału wykorzystywane są lasery z pionową wnęką rezonansową VCSEL. Pierwszy laser VCSEL zademonstrowano w roku 1979 [7]. Ich zaletą jest emisja z powierzchni lasera (inne typów laserów, np. Fabry-Perot czy DFB mają emisję krawędziową), co znacząco ułatwia łączenie ze światłowodem transmisyjnym, szczególnie o dużej średnicy rdzenia jak MMF OM4. Możliwe jest też testowanie laserów VCSEL bez pracochłonnego i kosztownego montażu poszczególnych komponentów. Pozwala to na selekcję komponentów na wczesnym etapie procesu produkcyjnego, co znacząco obniża koszt produkcji. Lasery VCSEL wymagają prądów zasilających rzędu kilku miliamperów, a emitowana moc optyczna jest rzędu kilku miliwat. Obecnie produkowane lasery VCSEL mają pasmo rzędu 25–30 GHz [8]. Pasmo fotodetektorów dla okna 850 nm jest rzędu 25–28 GHz.



» *Rys. 2. Przykładowe charakterystyki laserów VCSEL*
» *Fig. 2 Exemplary VCSEL characteristics*

Rys. 2. lewa kolumna pokazuje przykładową charakterystykę widma optycznego wielomodowego (MM) lasera VCSEL na długości fali 850 nm [9]. W widmie optycznym lasera widoczne jest wiele prążków, przez to całkowita szerokość widma sygnału jest rzędu kilku nanometrów. Taki sygnał optyczny podlega efektom dyspersji modowej i chro-

matycznej w światłowodzie wielomodowym, ograniczającym zasięg i/lub przepływność transmisji. Charakterystyka prądowo-mocowa (*Rys. 2.*) pokazuje, iż przy prądach zasilających 5-10 mA można uzyskać moce wyjściowe 0-5 dBm. Wykres pasma elektrycznego (uwzględnia on pasmo lasera VCSEL, pasmo światłowodu i pasmo odbiornika) pokazuje, iż zwiększanie długości światłowodu MMF znacząco ogranicza dostępne pasmo elektryczne, np. do 5 GHz dla długości MMF równej 600 m, co jest wynikiem działania efektu dyspersji modowej i chromatycznej. Dlatego też do realizacji datainterconnect o przepływności kilkudziesięciu Gb/s i zasięgu powyżej 100 metrów niezbędne są nowe techniki transmisji oraz komponenty opto-elektroniczne przezwyciężające ograniczenia pasma jak i efekty dyspersyjne.

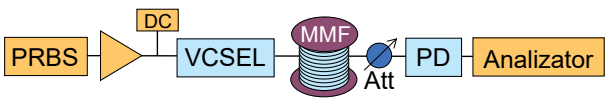
Jednym ze sposobów ograniczenia wpływu dyspersji modowej i chromatycznej jest wykorzystanie jednomodowych (SM) laserów VCSEL [10-11]. Lasery te charakteryzują się ograniczonym widmem optycznym jak i kompozycją modów. Ogranicza to wpływ efektów dyspersyjnych.

Oczywiście kluczowe jest tutaj zachowanie pasma elektrycznego systemu.

Rys. 2 prawa kolumna pokazuje najważniejsze charakterystyki przykładowego lasera SM VCSEL [8, 12]. Analizując widmo optyczne można zauważyć ograniczoną w stosunku do lasera MM VCSEL liczbę prążków w widmie optycznym. Ponadto moc prążków maleje o około 15 dB dla kolejnych składowych widma. Bezsprzecznie przyczynia się to do ograniczenia wpływu efektów dyspersyjnych. Charakterystyka prądowo-mocowa pokazuje mniejszy pobór prądu niż w przypadku lasera MM VCSEL. W przypadku lasera SM VCSEL typowe prądy zasilające wynoszą 2 mA. Wpływa to na moc generowanego sygnału, która jest tutaj ograniczona do -3 dBm. Przedstawiony wykres pasma elektrycznego dla SM VCSEL pokazuje, iż dla długości światłowodu do 600 m praktycznie nie występują ograniczenia pasma. Różni się to od lasera MM VCSEL, gdzie ograniczenie pasma było już widoczne dla długości światłowodu 100 m. Przedstawione charakterystyki laserów pokazują możliwość realizacji ultra-szybkiej transmisji o znacząco dłuższych odległościach dla laserów SM VCSEL niż MM VCSEL.

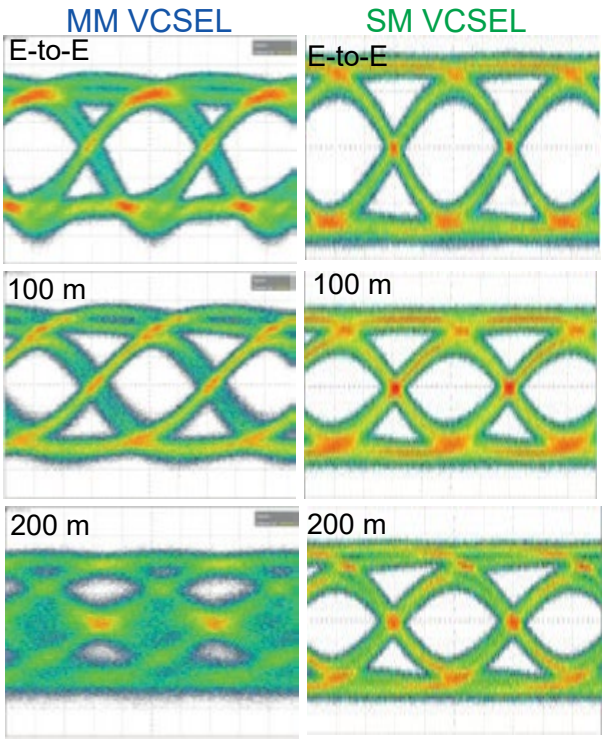
EKSPERYMENTY TRANSMISYJNE

Rys. 3. przedstawia schemat układu transmisyjnego. Jako źródło sygnału danych użyto generatora sekwencji pseudolosowej PRBS (*Pseudo Random Binary Sequence*). Sygnał PRBS był łączony z prądem o stałej wartości (DC), który ustalał punkt pracy lasera. Na wyjściu lasera generowany był sygnał zmodulowany amplitudowo, który był podawany do światłowodu MMF OM4 o różnej długości. Wartość mocy optycznej była ustalana w regulowanym tłumiku optycznym Att. Po konwersji na sygnał elektryczny w fotodetektorze PD, wyjściowy sygnał elektryczny był podawany do analizatora, np. oscyloskopu szerokopasmowego czy miernika bitowej stopy błędów.



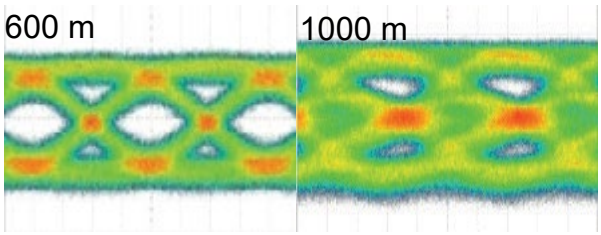
» *Rys. 3. Schemat układu transmisyjnego*
» *Fig.3. Transmission system scheme*

*Rys. 4. p*okazuje wykresy oka dla sygnałów na wyjściu systemu transmisyjnego dla różnych typów laserów jak i długości linii transmisyjnej. Oznaczenie E-to-E (*End-to-End*) oznacza bezpośrednie połączenie nadajnika z odbiornikiem optycznym za pomocą bardzo krótkiego (1-2 m) odcinka światłowodu. Jak widać w przypadku E-to-E wykres oka jest szeroko otwarty zarówno w dziedzinie czasu jak i amplitudy. Pokazuje to prawidłową generację sygnału o przepływności 25 Gb/s przez oba typy laserów. W przypadku transmisji na dystansie 100 m, oba sygnały wyglądają podobnie i wykres oka jest ponownie szeroko rozwarty, co wskazuje na prawidłową pracę systemu transmisyjnego na tym dystansie.



» *Rys. 4. Wykresy oka sygnałów o przepływności 25 Gb/s.*
» *Fig. 4. 25 Gb/s eye diagrams*

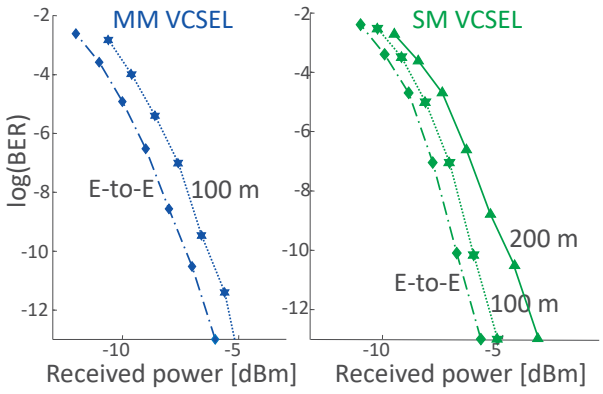
W przypadku transmisji na dystansie 200 m obserwujemy znaczące pogorszenie się jakości sygnału dla lasera MM VCSEL. Wykres oka jest zamknięty, co wskazuje na brak praktycznej możliwości realizacji transmisji na tym dystansie. Wykres oka dla SM VCSEL pozostaje szeroko otwarty. *Rys. 5.* pokazuje wykresy oka dla sygnałów 25 Gb/s w przypadku lasera SM VCSEL dla długości linii transmisyjnej 600 m i 1000 m. Dla długości światłowodu transmisyjnego równej 600 m wykres oka jest przymknięty, o ograniczonej rozwartości w dziedzinie czasu jak i amplitudy. Dla długości światłowodu 1000 m wykres oka jest kompletnie zamknięty.



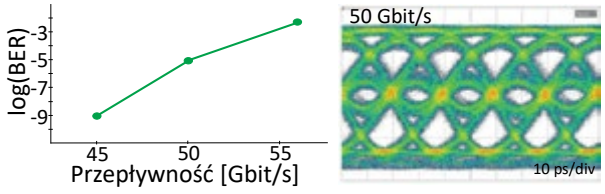
» *Rys. 5. Wykresy oka dla sygnałów 25 Gb/s SM VCSEL*
» *Fig. 5. SM VCSEL 25 Gb/s eye diagrams*

Analiza wykresów oka potwierdza obserwacje płynące z analizy charakterystyk pasma elektrycznego. Oba typy laserów mogą być wykorzystane w przypadku transmisji na dystansie do 100 m. Transmisja na dystansie 200 m i więcej może być tylko zrealizowana z wykorzystaniem SM VCSEL.

W dalszej kolejności przeprowadzono pomiary bitowej stopy błędów BER (*Bit Error Ratio*). Wyniki pomiarów BER przedstawiono na *Rys. 6.* W przypadku lasera MM VCSEL możliwa była bezbłędna (BER <1e-9) transmisja na dystansie 100 m. W przypadku transmisji z wykorzystaniem SM VCSEL, możliwa była bezbłędna transmisja na dystansie do 200 m. Można tutaj dodać, iż dla dystansu 300 m uzyskano najniższy BER nieznacznie powyżej 1e-9. Ważną obserwacją są ograniczone straty transmisji w przypadku SM VCSEL dla długości światłowodu 100 m, które są poniżej 1 dB w stosunku do przypadku E-to-E. Pozwala to na zredukowanie mocy nadajnika, co zmniejsza zużycie energii.



» *Rys. 6. Wyniki pomiarów BER*
» *Fig. 6. BER measurements*



» *Rys. 7. Wykres BER i oka dla ultra-szybkich przepływności dla lasera SM VCSEL.*
» *Fig. 7. BER measurements and eye diagrams for ultra-high speed SM VCSEL operation*

Pomiary dla ultra-szybkich przepływności w zakresie 45 Gb/s – 56 Gb/s dla lasera SM VCSEL w konfiguracji linii transmisyjnej E-to-E pokazały, iż jest możliwa bezbłędna transmisja z przepływnością 45 Gb/s. Dalsze zwiększanie przepływności wymaga zastosowania kodowania nadmiarowego, np. FEC (*Foreward Error Correction*). Z wykorzystaniem FEC możliwe jest uzyskanie bezbłędnej transmisji o przepływności 50 Gb/s.

Dalsze zwiększanie przepływności i/lub zasięgu transmisji wymaga zastosowania zaawansowanych formatów modulacji, np. PAM-4 (*Pulse-amplitude Modulation*), CAP (*Carrierless Amplitude Phase Modulation*) czy DMT (*Discrete Multitone Modulation*) oraz technik korekcji i przetwarzania sygnału. Rozwiązania te wymagają dodatkowych układów elektronicznych, które zwiększają powierzchnię układu, zużycie energii, opóźnienia i jego koszt. W pracy [13] zademonstrowano transmisję o przepływności 54 Gb/s (przepływność rzeczywista 50 Gb/s) na dystansie do 2,2 km MMF. Możliwe to było dzięki wykorzystaniu korekcji DFE (*Decision Feedback Equaliser)* i FFE (*Feed Forward Equaliser*). W pracy [14] zademonstrowano po raz pierwszy transmisję o przepływności 100 Gb/s z wykorzystaniem laserów VCSEL i światłowodów MMF. Było to możliwe dzięki zastosowaniu modulacji multi-CAP. Wykorzystując modulację DMT i nową generację laserów SM VCSEL o paśmie 30 GHz zrealizowano transmisję o przepływności 224 Gb/s [15].

WNIOSKI

Wzrost ruchu transmitowanego wewnętrz centrów danych pomiędzy poszczególnymi urządzeniami, takimi jak serwery, pamięci masowe, rutery etc jak i ich częściami składowymi wymaga rozwoju innowacyjnych systemów transmisyjnych. Kluczowe cechy systemów transmisyjnych w rozważanej aplikacji to przepływności rzędu kilkudziesięciu Gb/s, ograniczone opóźnienie oraz zminimalizowane zużycie energii. Systemy te mogą być z powodzeniem zrealizowane z wykorzystaniem laserów VCSEL, światłowodów wielomodowych i bezpośredniej modulacji. Ograniczenia medium transmisyjnego, w postaci efektów dyspersyjnych mogą znacząco ograniczać przepływność i zasięg transmisji tychże systemów. Przeprowadzone badania wykazały przewagę laserów SM VCSEL nad laserami MM VCSEL, a rozwój technik transmisyjnych z ich wykorzystaniem umożliwił znaczące zwiększanie zasięgu i przepływności systemów datainterconnect.

LITERATURA

[1] https://www.nobelprize.org/prizes/physics/2009/summary/
[2] https://www.ciena.com/products/wavelogic/wavelogic-6
[3] Schow, Clint L.. 2019. „Photonic technologies in the datacenter", Optical Fibre Conference, SC431.
[4] https://www.cables-solutions.com/whats-the-difference-om3-vs-om4.html
[5] https://www.ieee802.org/3/db/public/adhoc/presentations/abbott_3db_adhoc_01_080620.pdf
[6] https://www.prysmiangroup.com/en/en_fibre_multi-mode_bend-bright-tm-om5-mmf.html

[7] Soda H., Iga K., Kitahara C. i Suematsu Y.. 1979. "GaInAsP/InP surface emitting injection lasers". Jap. J.of Applied Physics: 18(12): 2329-2330.
[8] Ledentsov N.N., Makarov O., Shchukin V.A., Kalosha V.P., Ledentsov N., Chrochos Ł., Bou Sanayeh M. i Turkiewicz J.P.. 2022. "High speed VCSEL technology and applications". IEEE Journal of Lightwave Technology: 40(6):1749-1763.
[9] Turkiewicz J.P., Chorchos Ł., Shchukin V.A., Kalosha V.M., Kropp J-R., Augustin M. , Ledentsov Jr N. i Ledentsov N.N.. 2017. "High speed transmission with 850 nm SM and MM VCSELs" 19th International Conference on Transparent Optical Networks (ICTON): Mo.D3.5.
[10] Larsson, A. i Gustavsson, J.S.. 2013. "Single-Mode VCSEL". VCSELs. Springer Series in Optical Sciences, 166: 119-144.
[11] Unold H.J., Mahmoud S.W.Z., Jäger R., Kicherer M., Riedl M.C., Ebeling K.J.. 2000. "Improved single mode VCSEL performance by introducing a long cavity". IEEE Photonics Technology Letters: 12(8): 939-941.
[12] Ledentsov N.N., Shchukin V.A., Kalosha V.P., Ledentsov N. , Kropp J.-R., Agustin M., Chorchos Ł., Stępniak G., Turkiewicz J.P. i Shi J.-W.. 2018. "Anti–waveguiding vertical–cavity surface–emitting laser at 850 nm: From concept to advances in high–speed data transmission". Optics Express, 26(1): 445-453.
[13] Stepniak G., Lewandowski A., Kropp J.R., Ledentsov N.N., Shchukin V.A., Ledentsov N. Jr., Schaefer G., Agustin M., Turkiewicz J.P.. 2016. "54 Gbps OOK transmission using single mode VCSEL up to 2.2 km MMF". Ele. Letters 52(8): 633 – 635.
[14] Puerta Ramirez R., Agustin M., Chorchos Ł., Tański J., Kropp J.-R., Ledentsov Jr. N., Shchukin V.A., Ledentsov N.N., Henker R., Tafur Monroy I., Vegas Olmos J.J. i Turkiewicz J.P.. 2017. "107.5 Gb/s 850 nm multi- and single-mode VCSEL transmission over 10 and 100 m of multi-mode fiber", Optical Fiber Communication Conference paper Th5B.5,
[15] Ledentsov Jr. N., Chorchos Ł., Makarov O., Shchukin V.A., Kalosha V.P., Kropp J.-R., Turkiewicz J.P., Kottke C., Jungnickel V., Freund R., Ledentsov N.N. 2021. "Serial data transmission at 224 Gb/s applying directly modulated 850 and 910 nm VCSELs". Ele. Letters 57(19): 735-737.

Prenumerata cyfrowa

www.przegladtelekomunikacyjny.pl ***** www.sigma-not.pl

DOI: 10.15199/59.2023.4.4

Metamateriały, konfigurowalne matryce antenowe i komunikacja holograficzna
– wstępna analiza nowej koncepcji bezprzewodowej transmisji danych

Metamaterials, reconfigurable antenna arrays and holographic communication
– preliminary analysis of a new concept of wireless data transmission

dr hab. inż. ADRIAN KLIKS, prof. uczelni
Instytut Radiokomunikacji, Politechnika Poznańska
adrian.kliks@put.poznan.pl

STRESZCZENIE: W ostatnich kilku latach bardzo duże zainteresowanie wśród naukowców z całego świata zdobywa niezwykle oryginalna koncepcja komunikacji holograficznej. Specyfika tego podejścia z jednej strony niezwykle mocno odbiega od znanych i stosowanych obecnie rozwiązań, z drugiej stwarza bardzo duże możliwości rozwojowe w zakresie komunikacji bezprzewodowej. W artykule w sposób przeglądowy przedstawiono dwa rozwiązania technologiczne, które stały się przyczyną powstania idei komunikacji holograficznej. Jako pierwsze opisano możliwość wykorzystania tzw. meta-materiałów dla celów bezprzewodowej transmisji danych, jako drugie natomiast – zastosowanie rekonfigurowalnych powierzchni antenowych. W ostatniej części przedstawiono z kolei założenia idei komunikacji holograficznej, w której znane z holografii optycznej zasady tworzenia obrazów zostały przeniesione do pasma radiowego i do pewnego stopnia uogólnione.
SŁOWA KLUCZOWE: metamateriały, konfigurowalne matryce antenowe, komunikacja holograficzna, bezprzewodowa transmisja danych

ABSTRACT: In the last few years, a very original concept of holographic communication has gained a lot of interest among scientists from all over the world. The specificity of this approach, on the one hand, is very different from the known and currently used solutions, on the other hand, it creates great development opportunities in the field of wireless communication. The article provides an overview of two technological solutions that gave rise to the idea of holographic communication. First, the possibility of using the so-called meta materials for the purposes of wireless data transmission, and the second – the use of reconfigurable antenna surfaces. The last part presents the assumptions of the idea of holographic communication, in which the principles of creating images known from optical holography have been transferred to the radio band, and to some extend – generalized.
KEYWORDS: metamaterials, reconfigurable antenna arrays, holographic communication, new concept of wireless data transmission

Jednym z dominujących tematów w ostatnich latach w kontekście prowadzonych prac badawczych, wdrożeniowych i standaryzacyjnych w obszarze bezprzewodowych systemów telekomunikacyjnych było opracowanie rozwiązań dla tzw. piątej generacji (5G) sieci komórkowych [1]-[3]. Równocześnie w zakresie sieci lokalnych wiele prac podjętych było w celu upowszechnienia szóstej generacji (IEEE 802.11ax) oraz zdefiniowania nowych rozwiązań dla generacji siódmej (IEEE 802.11be) [4]-[8]. Równie dynamicznie rozwijają się inne rodzaje sieci, wystarczy wspomnieć tu o praktycznych implementacjach sieci Internetu Rzeczy [9]-[10], optycznej transmisji bezprzewodowej [11]-[13] czy komunikacji satelitarnej [14]-[15]. Należy zaznaczyć, że z perspektywy naukowej moment wdrażania i standaryzacji stanowi swojego rodzaju zakończenie pewnego etapu badań nad nowymi technikami czy nawet technologiami. Równolegle więc do wyżej wspomnianych działań społeczność naukowa pracuje nad rozwiązaniami, które być może znajdą zastosowanie w kolejnych (najbliższych, albo jeszcze późniejszych) generacjach systemów bezprzewodowych. Wystarczy wspomnieć, iż od kilku lat trwają intensywne dyskusje nad wymaganiami stawianymi szóstej generacji sieci komórkowych (6G), a także kolejnym generacjom sieci lokalnych [16]-[17]. W dyskusji tej wymienia się także różne przypadki zastosowań oraz rozwiązań technologicznych, które miałby zagwarantować wypełnienie wspomnianych wymagań.

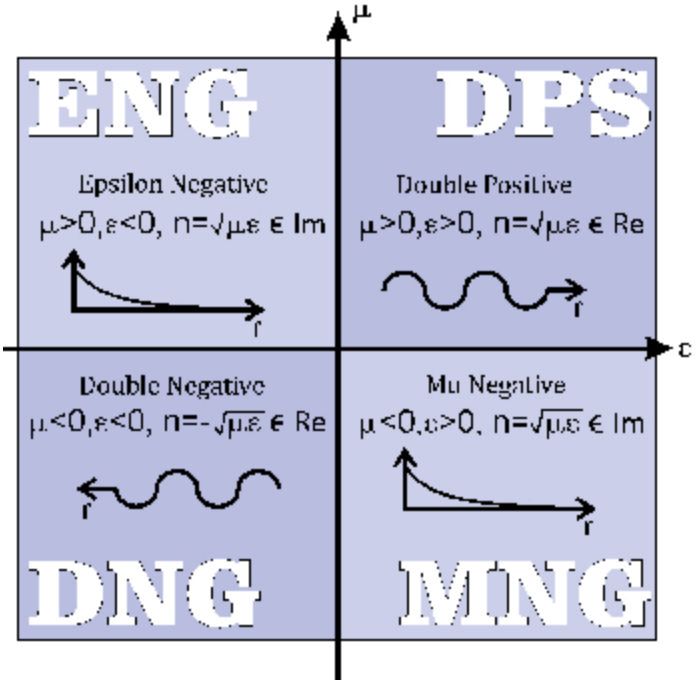
Wśród rozwiązań najczęściej pojawiających się w debacie naukowej należy wymienić m.in. powszechne użycie algorytmów sztucznej inteligencji (ang. *artificial intelligence*, AI), wykorzystanie możliwości wynikającej z komunikacji semantycznej (ang. *semantic communications*, SC) [18], zastosowanie programowalnych i otwartych sieci dostępu radiowego (*Open Radio Access Networks*, Open-RAN) [19] i IEEE 802.11 (Open-WiFi) [20], rozwiązań kwantowych [21]-[22], czy też techniki rejestrów rozproszonych dla komunikacji bezprzewodowej [23]. Skupiając się na warstwie fizycznej systemu telekomunikacyjnego należałoby wymienić w szczególności modulację OTFS (ang. *Orthogonal Time Frequency Space*) [24], techniki zintegrowanej komunikacji i nasłuchiwania widma (ang. *Integrated Communications and Sensing*, ICAS) [25], pełnej, jednoczesnej komunikacji dupleksowej w tym samym paśmie (ang. *In-Band Full Duplex Communications*, IBFD) [25], wykorzystanie pasma bardzo wysokich częstotliwości w tym fragmentu pasma światła widzialnego (ang. *Visible Light Communications*) [11], integrację transmisji radiowej i optycznej (ang. *Radio-Optical Communications*) [27]-[30], czy wreszcie rekonfigurowalne matryce antenowe RMA lub inteligentne powierzchnie rekonfigurowalne (ang. *Reconfigurable Intelligent Surfaces*, RIS) [31]-[33].

Propozycje wykorzystania matryc antenowych składających się z dużej liczby elementów w telekomunikacji bezprzewodowej są różnorodne. W klasycznym obecnie podejściu rozważa się w takiej sytuacji głównie techniki wieloantenowe (ang. *multiple input multiple outout*, MIMO lub w wersji z dużą liczbą anten – *massive MIMO*). Jednak w ostatnich latach ogromne zainteresowanie zdobyły wspomniane wcześniej rekonfigurowalne matryce antenowe, w których liczba elementów jest bardzo duża. Jednym z najprostszych i najczęściej spotykanych rozwiązań jest użycie RMA jako macierzy odbijających w sposób kontrolowany sygnał padający. Podejście takie pozwalałoby na np. takie przekierowanie sygnału padającego, aby po odbiciu został odebrany w miejscu, w którym bez macierzy RMA byłby on niedostępny lub zbyt słaby. W takim ujęciu RMA jest elementem pasywnym (tzn. bez aktywnego wzmocnienia sygnału). W innym ujęciu RMA mogłoby zostać wykorzystane do wytłumienia sygnału padającego w taki sposób, aby zminimalizować moc sygnału interferującego w zadanej lokalizacji. Możliwości użycia matryc są jednak znacznie większe i zostaną one omówione w kolejnych rozdziałach artykułu. Na szczególną uwagę zasługuje użycie tzw. meta-materiałów (meta-elementów) w procesie tworzenia powierzchni RMA. Intensywny rozwój technik materiałowych połączony z rozwojem elektroniki pozwolił na uzyskanie bardzo ciekawych efektów z perspektywy propagacji sygnałów. Bardzo intensywne prace w obszarze RMA a także wyniki badań nad zastosowaniem meta-materiałów dla komunikacji bezprzewodowej stały się przyczynkiem do powstania propozycji zupełnie nowego spojrzenia na transmisję sygnałów, nazywanej obecnie komunikacją holograficzną. Ponieważ nazwa może być myląca, należy podkreślić, że w swojej istocie komunikacja holograficzna nie jest to tylko zbiorem technik pozwalających na przesyłanie na odległość sygnału holograficznego. W niniejszym artykule skupiono się na przedstawieniu koncepcji komunikacji holograficznej, wskazując na jej postawy teoretyczne oraz wyzwania badawcze. W szczególny sposób uwzględniono możliwości zastosowania meta-materiałów oraz wykorzystania matryc RMA w celu praktycznej realizacji koncepcji komunikacji holograficznej [34]-[36].

Artykuł składa się z trzech głównych części. W pierwszej z nich (rozdział 2) przedstawiono ideę meta-elementów i meta-materiałów, skupiając się na możliwości ich użycia w propagacji sygnału. W trzecim rozdziale skupiono się na analizie rozwiązań z zakresu powierzchni RMA, a w czwartym – przedstawiono koncepcję komunikacji holograficznej, dyskutując jej podstawy teoretyczne oraz wyzwania. Wnioski płynące z przedstawionej w artykule analizy zawarto w rozdziale piątym.

META-MATERIAŁY

Propagacja fal elektromagnetycznych w różnego rodzaju środowiskach jest w bezpośredni sposób zależna od parametrów opisujących właściwości tego środowiska, w szczególności wymienić tu należy przenikalność elektryczną i magnetyczną. Pierwsza z nich (ang. *permittivity*) oznaczana jest często poprzez ε i jest wielkością łączącą indukcję pola elektrycznego $\vec{D}$ z natężeniem tego pola $\vec{E}$, . Druga z nich (ang. *permeability*) oznaczana jako μ i oznaczającej zdolność danego środowiska do zmiany indukcji pola magnetycznego $\vec{D} = \varepsilon\vec{E}$ w sytuacji zmiany natężenia tego pola $\vec{H}$, $\vec{B} = \mu\vec{H}$. Z kolei wszystkie cztery wielkości charakteryzujące pole elektromagnetyczne są ze sobą połączone poprzez podstawowe dla telekomunikacji prawa Maxwella, opisujące zasady propagacji fal radiowych.



» *Rys. 1. Podział materiałów ze względu na przyjmowane wartości przenikalności elektrycznej oraz magnetycznej*
» *Fig. 1. Division of materials due to the assumed values of permittivity and permeability*

Występujące naturalnie i stosowane obecnie materiały można sklasyfikować m.in. z perspektywy posiadanych przez nie właściwości przenikalności magnetycznej i elektrycznej. Wielkości te definiują także szeroko stosowany współczynnik załamania zgodnie ze wzorem $n = \pm\sqrt{\varepsilon\mu}$ (lub przedstawiany także jako $n^2 = \varepsilon\mu$). Warto zauważyć, że współczynnik załamania może przyjmować wartości zespolone, a składnik urojony związany jest z absorpcją ośrodka. W sytuacji, kiedy wartości obu przenikalności są dodatnie, możemy mówić o materiałach występujących naturalnie w środowisku czy o materiałach o typowych właściwościach. W literaturze anglojęzycznej pojawia się termin DPS (*double positive*). W ośrodku opisanym jako DPS propagacja sygnału odbywa się zgodnie z typowymi właściwościami wykorzystywanymi w transmisji sygnałów. W przypadku kiedy albo przenikalność elektryczna albo magnetyczna jest ujemna, współczynnik załamania staje się wartością urojoną, a transmisja sygnałów jest w ogólności niemożliwa ze względu na występujące ekspotencjalne tłumienie ośrodka. Materiały takie nazywane są często z angielskiego jako ENG (*epsilon negative*) i MNG (*mu negative*). Ostatnią z rozważanych tutaj sytuacji jest wystąpienie jednocześnie ujemnych wartości obu podatności – zarówno magnetycznej, jak i elektrycznej. W takiej sytuacji propagacja sygnału znów staje się możliwa, a materiały te nazywane są jako DNG (*double negative*)[1]. Materiały o takich cechach są możliwe do wyprodukowania przez człowieka, nie występują standardowo w środowisku [37], [38]. Zestawienie czterech klas materiałów przedstawiono na *Rys. 1*.

W przypadku ośrodka określonego jako DPS, zarówno wektory opisujące natężenie pola elektrycznego $\vec{E}$ oraz magnetycznego $\vec{H}$ oraz wektor falowy $\vec{k}$ są prawoskrętne, innymi słowy wektor Poytinga i wektor falowy są skierowane tak samo. Sytuacja jest inna dla materiałów DNG, bowiem w tym przypadku wektor falowy jest przeciwnie skierowany do wektora Poytinga, a wektory $\vec{k}$, $\vec{E}$ oraz $\vec{H}$ są lewoskrętne. Sytuacja ta przekłada się na zupełnie nietypowe zachowanie się fali propagującej się w tym ośrodku. Warto tu wspomnieć o odwróconym zjawisku Dopplera (przy zbliżaniu się do źródła fali długość fali będzie ulegała zwiększeniu), negatywnym zjawisku załamania (soczewki wypukłe i wklęsłe działałyby w odwrotny sposób) czy też o fakcie, że prędkość grupowa fali będzie przeciwna do prędkości fazowej. Wspomniane własności materiałów DNG wskazują na możliwe ciekawe zastosowania w różnych dziedzinach życia, w tym w szczególności w telekomunikacji. Dla przykładu możliwość wpływania na zjawisko refrakcji mogłaby być z powodzeniem wykorzystania do kontrolowanego sterowania wiązką sygnału [40].

Jednakże możliwość wytworzenia materiałów DNG nie wyczerpuje samo w sobie możliwości oferowanych przez tzw. meta-materiały. Słowo meta-materiał wskazuje na materiał, którego cechy są „poza" tymi występującymi typowo w naturze. W ogólności więc, jako meta-materiał można traktować każdą sztucznie wytworzoną i powtarzalną strukturę (materiał), a powtarzający się element podstawowy w takiej strukturze nazywa się również także *meta-atomem*. Sztucznie wytworzony meta-materiał nie musi być więc jednolity, może składać się z powtarzalnych struktur, które traktowane razem będą składały się na powierzchnię meta-materiału. Z perspektywy telekomunikacji warto zwrócić uwagę na ciekawy przykład budowy meta-materiału składającego się z powtarzających się rezonatorów magnetycznych (*split-ring resonators*, SRR, prostych cewek odpowiedzialnych za reagowanie na zmiany pola magnetycznego) oraz małych dipoli (odpowiedzialnych za reagowanie na zmiany pola elektrycznego) [39]. Korzystając z tak zdefiniowanego *meta-atomu* udało się skonstruować trójwymiarowe struktury (*omni box*) pracujące na częstotliwościach rzędu kilku GHz. Zwróćmy uwagę na fakt, że wytwarzanie meta-materiałów w taki sposób otwiera nowe możliwości kontrolowania właściwości elektromagnetycznych tak wytworzonej struktury.

W tym kontekście bardzo często wskazuje się, że odstęp pomiędzy sąsiadującymi elementami w strukturze powinien być z założenia mniejszy od połowy ($\lambda/_2$) lub nawet czwartej części ($\lambda/_4$) długości fali, jest to więc założenie inne niż stosowane powszechnie w kontekście technik wieloantenowych. W pracy [40] wskazano, że właściwości meta-materiałowych struktur mogą być zmieniane na kilka sposobów. Pierwszy z nich to modyfikacja mechaniczna, gdzie zmienia się bezpośrednio odległość pomiędzy *meta-atomami* wpływając przez to na wypadkowe właściwości elektro-magnetyczne całej struktury. Jest to jednak metoda relatywnie kosztowna i dość złożona. Inne metoda modyfikacji polega na zmianie cech elektro-magnetycznych całego układu poprzez zmiany składowych elementów elektronicznych (wykorzystuje się tutaj np. diody PIN, waraktory, układy MEMS). Trzecia opcja zakłada zmianę właściwości struktury poprzez odpowiednie oddziaływania temperaturowe, czwarta natomiast wykorzystuje różne możliwości oddziaływania chemicznego na użyte substancje składowe struktury (dość często wymienia się tutaj grafen).

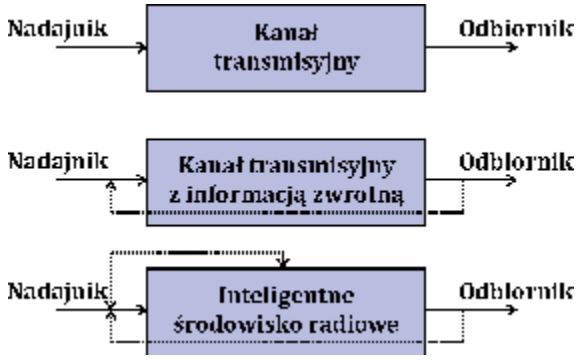
Z kolei z perspektywy wytwarzania meta-materiałów w pracy [40] wymieniono kilka metod wartych uwagi m.in. fotolitografia (stosowana dla jedno lub wielowarstwowych meta-materiałów pracujących w zakresie długości fali od 30 nm do 3 mm), miękka fotolitografia, elektronolitografia (litografia wykorzystująca wiązkę elektronów) czy też SML (*shadow mask litography*), zwrócono także uwagę na techniki trójwymiarowe 3D.

Warto podkreślić, że rozwój metamateriałów stwarza szerokie możliwości aplikacyjne w różnych dziedzinach życia dzięki wykorzystaniu zarówno struktur określanych jako DNG, jak i tych składających się powtarzających się elementów składowych (*meta-atomów*). Wskazane wcześniej możliwości kontroli właściwości metamateriałów stały się inspiracją do powstania koncepcji materiałów programowalnych lub materiałów definiowanych programowo (*software defined materials* – SDM, *software programmable metasurfaces* – SPM) [41]-[43]. Analogicznie do idei radia czy sieci definiowanych programowo, zakłada się tutaj możli-

[1] W zasadzie w anglojęzycznej literaturze przedmiotu można spotkać różne nazwy stosowane dla określenia materiałów o ujemnych wartościach przenikalności magnetycznej i elektrycznej. Są to m.in. NGV - *negative group velocity,* NIM - *negative index materials,* DNM - *double negative materials,* LHM - *left-handed materials*. W języku polskim można spotkać nazwy np. materiały z ujemną refrakcją, materiały z ujemnym współczynnikiem załamania, podwójnie ujemne materiały, materiały z falą wsteczną czy wreszcie materiały z ujemną prędkością fazową.

wość wpływania na właściwości powierzchni wykonanej z meta-materiału z poziomu oprogramowania. Z kolei autorzy pracy [44] zaproponowali, aby najmniejszy powtarzalny element składowy (*meta-atom*) – nazywany super-komórką (*supercell*) – składał się z bardzo szerokiego wachlarza sub-elementów, gdzie każdy z nich jest odpowiedzialny za inne funkcje elementarne (np. pobieranie energii z otoczenia, dystrybucję energii w ramach elementu składowego, za komunikację w pasmie wysokich częstotliwości, komunikację w pasmie podczerwieni itp.).

Podsumowując powyższą jednak dość pobieżną analizę warto zwrócić uwagę na dynamikę rozwoju koncepcji meta-materiałów, zwłaszcza z perspektywy ich zastosowania dla telekomunikacji. Pojawienie się materiałów z ujemną refrakcją z jednej strony oraz struktur fabrykowanych z wykorzystaniem powtarzalnych elementów składowych (*meta-atomów*) z drugiej pokazuje, że materiały te można potencjalnie wykorzystać do istotnego zmieniania warunków propagacyjnych. Materiały takie można użyć zarówno po stronie nadawczej i odbiorczej, jak i umieścić je w postaci matryc RMA bezpośrednio w środowisku propagacyjnym, wpływając na jego cechy. Należy jednak podkreślić, że wykorzystanie meta-materiałów w przypadku RMA jest tylko jedną z możliwości ich zastosowania; matryce RMA mogą bowiem być rozmieszczone w środowisku jako fragment systemu telekomunikacyjnego niekoniecznie opierając się na meta-materiałach (zwłaszcza tych określanych jako DNG).



» *Rys. 2. Ilustracja koncepcji inteligentnego środowiska radiowego (smart radio environment, SRE)*
» *Fig. 2. Illustration of the Smart Radio Environment, SRE, concept*

INTELIGENTNE MATRYCE ANTENOWE

W kolejnych generacjach systemów bezprzewodowych (komórkowych, lokalnych) coraz większą rolę odgrywały techniki wieloantenowe MIMO, w których liczba elementów antenowych systematycznie wzrastała. Początkowo były to dwie lub kilka anten, później liczba ta wrosła do kilkunastu lub nawet kilkudziesięciu, aby ostatecznie mówić o technice massive-MIMO, gdzie liczba elementów antenowych jest co najmniej równa 128. W tym czasie równolegle rozpoczęła się dyskusja naukowa o możliwościach wykorzystania matryc antenowych RMA jako do pewnego stopnia niezależnych i autonomicznych elementów całego systemu telekomunikacyjnego. Wykorzystanie matryc antenowych RMA zazwyczaj łączone jest z koncepcją określaną z języka angielskiego jako inteligentne środowisko radiowe (*smart radio environment*, SRE). Autorzy pracy [45] zdefiniowali dwa bardzo istotne pytania, które w dokładny sposób odzwierciedlają założenia SRE:

- Jakie perspektywy stworzyłaby możliwość monitorowania (tj. dokonywania obserwacji i raportowania pomiarów) środowiska bezprzewodowego poprzez rozmieszczenie urządzeń wykorzystujących do tego celu tylko i wyłącznie sygnały już w tym środowisku istniejące, w dodatku bez konieczności użycia baterii?
- Co dałaby możliwość wpływania na środowisko propagacyjne (za pomocą kontrolera sterowanego zdalnie w sposób programowy i wyposażonego w możliwości predykcji) w celu np. zwiększenia przepływności bez istotnego zwiększania zużycia energii?

Pytania te mogą się w pierwszej chwili wydawać bardzo abstrakcyjne i dalekie od praktycznego zastosowania, jednak przy przyjęciu pewnych dodatkowych założeń pokazują ciekawe obszary badawcze z perspektywy propagacji fal radiowych, transmisji danych i modelownia systemów telekomunikacyjnych. Wzorując się na [45] na *Rys. 2* przedstawiono klasyczne podejście do transmisji danych (gdzie zarówno po stronie odbiorczej i nadawczej podejmowane są działania w celu minimalizacji negatywnego wpływu kanału propagacyjnego), jak i podejście SRE (gdzie dodatkowo pojawia się opcja oddziaływania na środowisko propagacyjne).

Jedną z proponowanych technik wpływania na parametry środowiskowe w przypadku propagacji fal elektro-magnetycznych jest właśnie wykorzystanie możliwości stwarzanych przez meta-powierzchnie (czyli powierzchnie stworzone z wykorzystaniem meta-materiałów). W prostszym ujęciu, meta-powierzchnie mogą być zastąpione przez inteligentne powierzchnie rekonfigurowalne RIS lub też rekonfigurowalne matryce antenowe RMA. W tradycyjnym ujęciu RIS lub RMA składa się z wielu gęsto upakowanych i konfigurowalnych elementów antenowych, których odległość między nimi jest większa niż czwarta część lub połowa długości fali. Dodatkowo każdy z elementów antenowych jest sterowany np. za pomocą przesuwnika fazowego (ang. *phase shifter*) [37]. Dzięki takiemu podejściu RMA jest traktowany jako element pasywny (tzn. bez aktywnego wzmocnienia sygnału). Jak już wcześniej wspomniano, RIS czy RMA mogłoby zostać wykorzystane do wytłumienia sygnału padającego w celu minimalizacji interferencji w zadanej lokalizacji.

W literaturze naukowej można odnaleźć setki bardzo zaawansowanych publikacji związanych z zastosowaniem matryc RIS/RMA w telekomunikacji. Możliwości stwarzane przez powierzchnie RIS/RMA są na tyle obiecujące, że Europejski Instytut Norm Telekomunikacyjnych (*European Telecommunications Standards Institute,* ETSI) powołał specjalną grupę roboczą zajmującą się tym zagadnieniem (ang. *Industry Specification Group* (ISG) *on Reconfigurable Intelligent Surfaces* (RIS)). Jako RIS rozumie się tutaj węzeł sieci

bezprzewodowej wyposażony w specjalną powierzchnię zbudowaną z elementów antenowych lub meta-materiałów, posiadającą specyficzne właściwości odbijające, refrakcyjne i absorpcyjne, które mogą być zmienione w zależności od panujących warunków środowiskowych. Grupa ISG-RIS zajmuje się wielorakimi zagadnieniami związanymi z użyciem powierzchni RIS do komunikacji bezprzewodowej, w tym: definicję przypadków użycia i scenariuszy rozmieszczenia, określenie wyzwań i możliwości technologicznych, zastosowanie do lokalizacji i nasłuchiwania widma, do zwiększenia bezpieczeństwa i prywatności w komunikacji, do minimalizacji ekspozycji na promieniowanie elektro-magnetyczne w wybranej lokalizacji itp. Ponadto członkowie grupy starają się opracować scenariusz referencyjny przeznaczony dla kompletnego testowania systemów *end-to-end*, w których wykorzystuje się węzły RIS. W kwietniu i czerwcu 2023 r. w grupa w ramach ETSI opublikowała dwa raporty dotyczące możliwości zastosowania płaszczyzn RIS [46]-[47]. Warto w tym miejscu zacytować, jak węzeł RIS jest interpretowany przez ISG RIS:

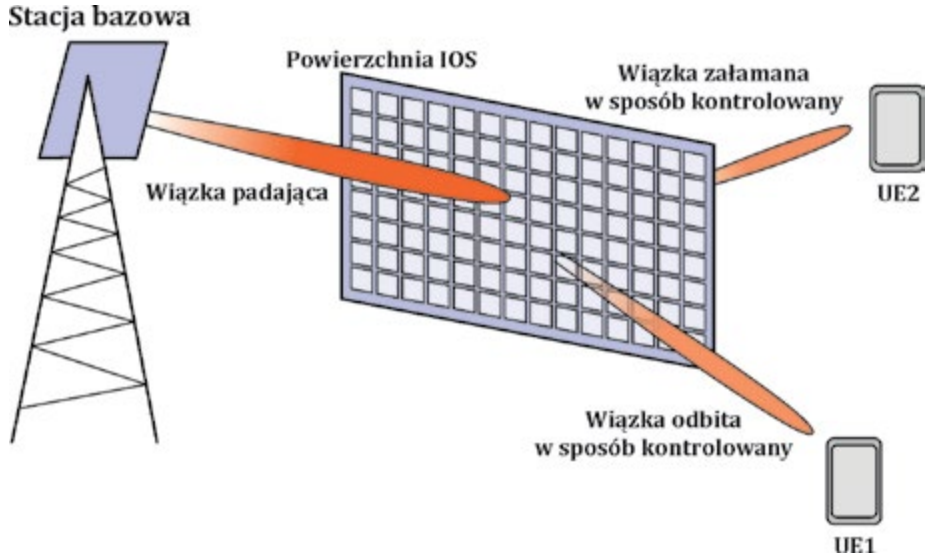
„RIS can be implemented using mostly passive components without requiring high-cost active components such as power amplifiers, resulting in low implementation cost and energy consumption. This allows easy and flexible deployment of RIS, with the possibility of RIS taking any shape and to be integrated onto objects (e.g. walls, buildings, lamp posts, etc.). RIS are supposed to run as nearly-passive devices and hence are unlikely to increase exposure to EMF, and in fact, they can potentially be used to reduce EM pollution in legacy deployments. These associated characteristics suggest RIS may be considered as a sustainable environmentally friendly technology solution. RIS may have different structures with considerations of cost, form factor, design and integration." [Cytat za [46]]

Macierze RMA można podzielić na aktywne i pasywne ze względu na technikę wytwarzania. W tym pierwszym przypadku mówimy o macierzach, w skład których wchodzą energochłonne elementy wysokoczęstotliwościowe oraz jednostki odpowiadające za przetwarzanie sygnałów. Z kolei powierzchnie pasywne działają jakby lustro metaliczne, które może być odpowiednio przeprogramowane w zależności od sposobu, w jaki ma zmienić właściwości fali odbitej; typowo układy te nie potrzebują dodatkowego źródła zasilania. Warto zauważyć, że technika wytwarzania macierzy RMA nie wpływa na jej sposób pracy – macierz oznaczona jako aktywna może działać w trybie bez zapotrzebowania na dodatkową energię.

Skupiając się na możliwych trybach pracy, wskazano na następujące warianty [47]:

- odbijające powierzchnie RMA pozwalające na zmianę kąta odbicia (ang. *reflective surfaces*),
- refrakcyjne (załamujące) powierzchnie RMA pozwalające na zmianę kąta załamania (ang. *refractive surfaces*),
- powierzchnie RMA posiadające obie powyższe własności, które mogą być wykorzystane jednocześnie (ang. *joint reflective and refractive surfaces*),
- absorbujące powierzchnie RMA używane w celu minimalizacji efektu rozpraszania (ang. *absorbing surfaces*),
- powierzchnie RMA transmisyjne lub informacyjne, posiadające możliwość zakodowania danych oraz transmisji radiowej (ang. *transmitting and information surfaces*)
- powierzchnie odbijające i refrakcyjne pozwalające jednocześnie na modulację fali odbitej i załamanej sygnałem z danymi (ang. *surface for ambient backscattering*)
- powierzchnie zwiększające rozpraszanie sygnałów w danym obszarze (ang, *surfaces for tuned randomness*),
- powierzchnie pozwalające na jednoczesne odbicie fali i wykrywanie obecności obiektów (ang. *communication and sensing surfaces*).

W literaturze odnaleźć można wiele ciekawych prac opisujących zarówno aspekty teoretyczne transmisji sygnałów w systemie wykorzystującym RMA, jak i praktyczne ich implementacje. Autorzy bardzo ciekawej pracy [48] zaproponowali wykorzystanie powierzchni IOS (ang. *Intelligent Omni Surfaces*), które posiadają zarówno możliwości odbijające i refrakcyjne zgodnie z *Rys. 3.*



» *Rys. 3. Powierzchnia IOS umożliwiająca zarówno kontrolowane odbicie jak i załamanie wiązki za [48]*
» *Fig. 3. IOS surface enabling both controlled reflection and refraction of the beam (following [48])*

Działanie takich macierzy umożliwiałoby więc jednoczesne skierowanie fali odbitej oraz załamanej w pożądanych kierunkach. Właściwości odbijające zaproponowanej matrycy IOS zostały osiągnięte dzięki użyciu modułów składających się z diod PIN, promienników (reflective patch) oraz obszarów będących masą układu, a także metalowej osłony minimalizującej zjawisko załamania. W przypadku zjawiska refrakcji posłużono się dwoma takimi modułami (ze zmodyfikowanym promiennikiem) złączonymi ze sobą i połączonymi ze sobą szczeliną (*via hole*). Połączenie tych dwóch rozwiązań ze sobą jednocześnie dało Autorom możliwość osiągnięcia jednoczesnego kontrolowanego zjawiska odbicia i załamania. W artykule przedstawiono równoważny model elektryczny modułu IOS i opisano macierze przejść dla takiego modelu. Dodatkowo zaproponowano opis matematyczny modelu propagacyjnego, w którym założono obecność elementów IOS pozwalających na formowanie wiązek odbitej oraz załamanej. Co ciekawe, przedstawiono także wyniki eksperymentów pomiarowych z wykorzystaniem kontrolowanych macierzy IOS. Na szczególną uwagę zasługują opisane eksperymenty z użyciem macierzy IOS jako materiału na szyby w oknach umożliwiających sterowanie sygnałem padającym na tę szybę.

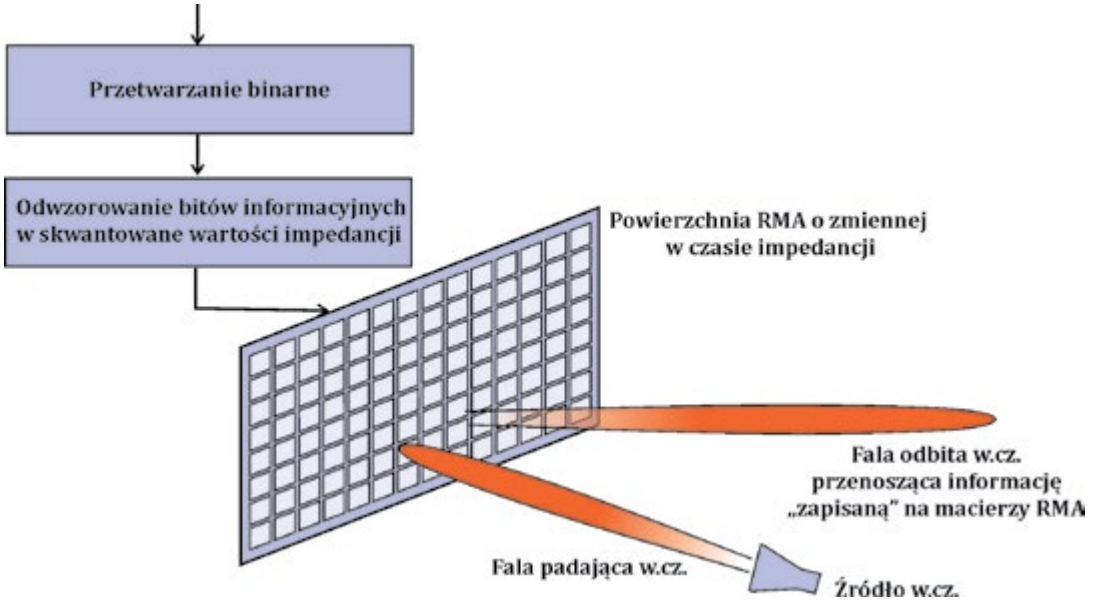
Ciekawe wyniki eksperymentu pokazano w pracy [49], gdzie Autorzy za pomocą matrycy RMA umieszczonej na zewnątrz budynku byli w stanie istotnie zwiększyć moc sygnału obserwowanego w badanym obszarze. W szczególności rozważono sytuację, w której w wybranej lokalizacji za budynkiem jakość sygnału 5G mierzona za pomocą miary RSRP (ang. *Reference Signal Receive Power*) była niewystarczająca – właśnie z powodu dużego tłumienia spowodowanego przeszkodą (rogiem budynku i drzewami). Zmierzona wartość RSRP bez użycia powierzchni RMA była w badanym obszarze mniejsza niż -110 dBm lub nawet niż -120 dBm. Dzięki umieszczeniu na przenośnym maszcie matrycy RIS, w badanym obszarze wartość RSRP wzrosła do zakresu od -90 dBm do -105 dBm.

Możliwości oferowane przez matryce RMA są jednak bardzo szerokie. Bardzo interesujące podejście użycia matryc RMA pokazano w pracy [50], gdzie w konstrukcji elementu podstawowego matrycy (meta atomu) wykorzystano waraktory. Dzięki opisowi matematycznemu przedstawiono równoważną impedancję poszczególnego modułu, jak i powstałej płaszczyzny. Założono przy tym, że dzięki odpowiednim zmianom napięcia, impedancja poszczególnego układu podstawowego może zmieniać się w czasie (w pracy przyjęto, że może przyjąć jedną z czterech wartości). Innymi słowy impedancja ta stała się matematycznie funkcją czasu i można ją traktować jako sygnał informacyjny (bity informacyjne są odwzorowywane w odpowiednie wartości impedancji wypadkowej). W proponowanym rozwiązaniu Autorzy oświetlają tak stworzoną matrycę za pomocą wysokoczęstotliwościowej fali padającej, uzyskując przez to przeniesienie sygnału informacyjnego (impedancji matrycy zapisanej w funkcji czasu) do pasma wysokich częstotliwości bez wykorzystania mieszacza, filtrów oraz szerokopasmowego wzmacniacza mocy. Eksperyment sprzętowy wykonano z wykorzystaniem macierzy FPGA oraz płyt radia definiowanego programowo SDR (ang. Software Defined Radio). Schemat poglądowy proponowanego rozwiązania zilustrowano na *rys. 4.*

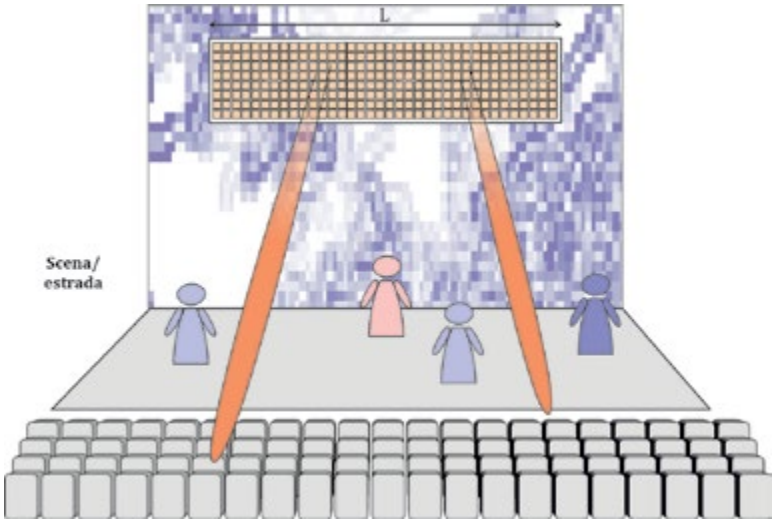
Przytoczone powyżej przykłady pokazują bardzo szerokie spektrum zastosowania macierzy RMA do transmisji sygnałów. W tym ujęciu klasyczne rozumienie płaszczyzn RIS/RMA jako elementy tylko odbijające i pasywne okazuje się być tylko jednym z najprostszych wariantów ich wykorzystania w praktyce. Warto tutaj nadmienić, że literatura przedmiotu dostępna w bazach elektronicznych i dotycząca różnych aspektów macierzy RMA jest niezwykle bogata.

KOMUNIKACJA HOLOGRAFICZNA

Przedstawiony ostatni przykład zastosowania macierzy RIS/RMA do transmisji sygnałów, w którym wyeliminowano z toru nadawczego układy mieszacza oraz filtry, wskazuje na nowy kierunek prac w obszarze telekomunikacji bez-



» *Rys. 4. Ilustracja koncepcji przesyłania informacji zapisanej w postaci zmiennej impedancji płaszczyzny RMA za [50],*
» *Fig. 4. An illustration of the information transfer concept written in terms of the RMA's plane's variable impedance (following [50])*

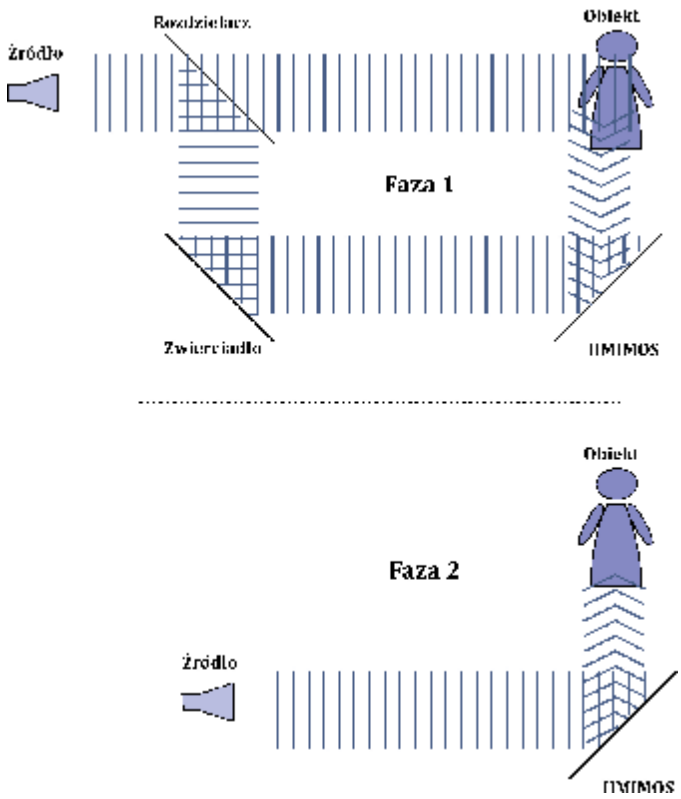


» *Rys. 5. Przykład zastosowania płaszczyzny RMA o dużych wymiarach skutkujący możliwością wystąpienia transmisji danych w polu bliskim*
» *Fig. 5. An example of using a large dimension RMA resulting in the possibility of the near field data transmission*

przewodowej jakim jest tzw. komunikacja holograficzna. W tym miejscu należy podkreślić dwa spostrzeżenia. Pierwsze, komunikacja holograficzna w ujęciu przedstawianym w tym artykule nie jest transmisją sygnałów holograficznych (np. obrazów holograficznych, strumienia danych wykorzystywanych w okularach rozszerzonej rzeczywistości itd.) przez sieć bezprzewodową. Drugie, koncepcja komunikacji holograficznej jest na tyle nowym obszarem badawczym, że podczas studiów literaturowych można natrafić na zupełnie odmienne sugestie jej rozwoju proponowane przez różnych autorów, łącznie z różnicami dotyczącymi stosowanego słownictwa czy podstawowymi zjawiskami, które w przypadku komunikacji holograficznej powinno się rozważyć. W tym rozdziale przybliżymy podstawowe pojęcia i obserwacje związane z omawianym zagadnieniem.

W przypadku zastosowania macierzy RMA składającej się z wielu elementów podstawowych, zupełnie racjonalnym założeniem jest przyjęcie rozmiaru takiej macierzy L liczonej w centymetrach albo nawet metrach. Dodatkowo można by założyć, że odległość pomiędzy nimi jest zdecydowanie mniejsza od połowy czy nawet czwartej części długości fali padającej, tak że pomiędzy elementami tej macierzy dochodzi do wzajemnych i istotnych oddziaływań elektro-magnetycznych. W konsekwencji efektywna apertura takiego układu będzie właśnie wyrażona także w centymetrach czy metrach [51]. Na *Rys. 5* przedstawiono jedno z przykładowych zastosowań takiej macierzy jako dużego elementu umieszczonego nad sceną lub estradą koncertową. Dodajmy, że zgodnie z analizą przedstawioną w poprzednich rozdziałach, macierze RMA mogą być wykonane z użyciem meta-materiałów, których nawet obecnie grubość jest bardzo mała[2].

Zwróćmy uwagę, jakie w takiej sytuacji występują wartości granicy pomiędzy polem dalekim (ang. *far field*) a polem bliskim (ang. *near field*). Korzystając z prostych zależności [51], dla macierzy RMA o długości L = 10 cm i dla częstotliwości około 10 GHz, granica między tymi strefami wynosi około kilku milimetrów. Jednak w przypadku zwiększenia rozmiaru macierzy do 100 cm, granica ta pojawia się w odległości prawie 100 m od środka macierzy. W praktyce oznacza to, że dla większych rozmiarów macierzy RMA (a takie są przecież rozważane), komunikacja będzie odbywała się bezpośrednio w polu bliskim. Na *Rys. 5* zobrazowane jest to za pomocą dwóch wiązek, które w zasadzie mają charakter tylko ilustracyjny, bowiem transmisja w zależności od długości L może odbywać się w polu bliskim. To zaś pociąga za sobą różnego rodzaju konsekwencje i zmusza do poszukiwania potencjalnie nowych rozwiązań z zakresu transmisji sygnałów uwzględniających szeroki wachlarz zjawisk pojawiających się w polu bliskim. W tym kontekście Autorzy wspomnianej pracy [51] proponują używanie terminu komunikacja holograficzna, aby podkreślić możliwości całościowego wykorzystania różnorodnych zjawisk pojawiających się w różnych odległościach od źródła (w polu bliskim i dalekim) i określenia limitów teoretycznych wynikających z własności propagacji fali elektro-magnetycznej w tych obszarach. Wskazują oni, że słowo holograficzny, wywodzące się z starogreckich słów *ὅλος* (*holos, all*), *and* *γραφή*



» *Rys. 6. Ilustracja dwóch faz komunikacji holograficznej*
» *Fig. 6. Two phases of the holographic communications*

[2] Obrazowo można powiedzieć, że takie macierze RMA mogłyby być traktowane „naklejki" na ścianę, okno czy inne przedmioty, mogłyby być także traktowane jako rodzaj faktury ściany.

(*graphé, writing*), oznacza literalnie „opisujący wszystko". Jest to jednak jedna z interpretacji dostępna w literaturze.

W bardzo obszernej pracy [52] (udostępnionej w trakcie tworzenia tego artykułu w repozytorium arXiv), Autorzy zaproponowali następującą definicję bezprzewodowej komunikacji holograficznej:

„Holographic wireless communication—it is the physical process of realistically and completely restoring the three dimensional (3D) target scene transmitted by the transceiving ends with the help of new holographic antenna technology and wireless EM signal technology, and at the same time realize 3D remote dynamic interactions with people, objects and their surrounding environment." [cytat za [52]]

W definicji tej warto zwrócić uwagę na kilka aspektów, po pierwsze wspomniana jest przestrzeń trójwymiarowa przenosząca informację, po drugie – zakłada się wykorzystanie wszelkich zjawisk obecnych w propagacji fal elektro magnetycznych, po trzecie zaś – wspomniana jest koncepcja anten holograficznych. W pracy [53] zaproponowano używanie zwrotu Holographic MIMO Surfaces (HMIMOS), podkreślając fakt wykorzystania typowych zjawisk i rozwiązań stosowanych w tradycyjnej holografii optycznej do transmisji w paśmie fal radiowych. W tym przypadku macierz MIMO (także wykonana z meta-materiałów, więc w istocie również powierzchnia RMA) pełni rolę powierzchni odpowiedzialnej za wytworzenie obrazu holograficznego w pasmie fal radiowych. Autorzy wzorując się na dwuetapowym procesie wytwarzania obrazów w holografii optycznej, proponują podobne podejście w odniesieniu do komunikacji bezprzewodowej. W szczególności w pierwszym etapie (nazywanym *holographic training*) zakłada się tutaj występowanie fali koherentnej, która rozdzielona na dwie składowe (falę padającą na obiekt oraz falę referencyjną) tworzą razem zestaw minimów i maksimów interferencji obserwowany i zapamiętany na płaszczyźnie HMIMOS. W drugiej fazie (nazwanej jako *holographic communications*) fala referencyjna odbita od płaszczyzny HMIMOS powoduje powstanie pożądanego obrazu holograficznego w odbiorniku. Zostało to zobrazowane na *Rys. 6.* Innymi słowy, w tym ujęciu komunikacja holograficzna stanowi swoiste przeniesienie rozwiązań znanych z holografii optycznej do obszaru fal radiowych.

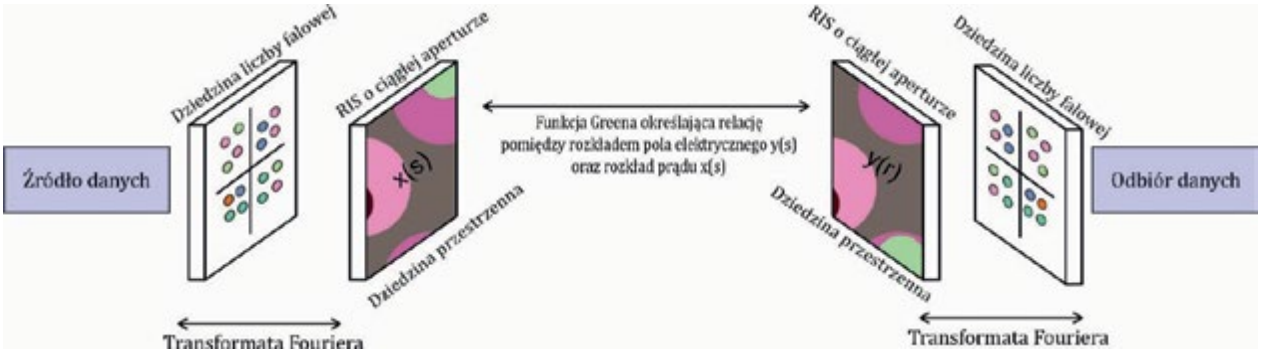
W konsekwencji takiego podejścia i z perspektywy analizy przesyłanej informacji, jak przedstawiono w [54], przestrzeń informacyjna staje się trójwymiarowa, bowiem odbiornik analizuje zachowania się wspomnianych ekstremów interferencji. Bardzo szczegółowy opis transmisji i odbiory takiego rodzaju sygnału przedstawiono m.in. w [52] oraz [55].

Wzorując się na [51] oraz [52] można zauważyć, że tradycyjne rozwiązania stosowane obecnie w telekomunikacji bezprzewodowej (dostosowane do propagacji fali w polu dalekim) zostają zastąpione rozwiązaniami typowymi dla pola bliskiego. W szczególności wykorzystując własności holografii opisane powyżej, rozważa się koncepcję, w której po stronie nadawczej stosuje się zbiór funkcji dwuwymiarowych (opisujących np. rozkład prądów czy ogólniej pola elektro-magnetycznego na powierzchni RMA, traktując tę powierzchnię jakby jako obraz), których rozkład na powierzchni RMA zależy od bitów informacyjnych. Obrazy takie są odtwarzane po stronie odbiorczej i analizowane z wykorzystaniem innego, dopasowanego zbioru funkcji odbiorczych. W [52] z kolej zaproponowano wykorzystanie tensorów Green'a do przesyłania informacji w przypadku kanału LOS (*Line of sight*), zaś dla kanału NLOS (Non-LOS) – zastosowanie rozwinięcia sygnału odbieranego w dwuwymiarowy szereg Fouriera (odzwierciedlający falę płaską na powierzchni). Rozwiązanie dla kanału LOS/NLOS takie przedstawiono graficznie na *Rys. 7.*

Bardzo ciekawym rozwiązaniem, inspirującym wielu badaczy, jest wykorzystanie orbitalnego momentu pędu światła (ang. *orbital angular momentum*, OAM) znanego z sieci optycznych właśnie do komunikacji holograficznej z wykorzystaniem macierzy RMA [56].

Już pobieżna analiza rozwiązań stosowanych w komunikacji holograficznej wskazuje, że obserwujemy bardzo istotny rozwój nowej koncepcji bezprzewodowej transmisji sygnałów. W szczególności można zauważyć następujące różnice w komunikacji holograficznej względem rozwiązań tradycyjnych:

- w przypadku komunikacji holograficznej transmisja odbywać się może zarówno w polu dalekim, jak i bliskim, a ze względu na duże rozmiary powierzchni RMA w praktyce dominujący będą będzie transmisja w polu bliskim;
- pociąga to za sobą fakt, że czoło fali nie będzie mogło być opisywane i traktowane jako płaskie,
- z perspektywy macierzy RMA w komunikacji holograficznej zakłada się, że odległość pomiędzy elementami składowymi jest znacznie mniejsza od połowy długości fali; jest



» *Rys. 7. Przekazywanie danych w systemie z matrycami RIS (HMIMOS) dla kanału LOS oraz NLOS*
» *Fig. 7. Data transmission in the system equipped with HMIMOs for LOS and NLOs channel*

to zupełne przeciwieństwo typowych założeń stosowanych w technice wieloantenowej, gdzie minimalna odległość pomiędzy elementami aktywnymi macierzy MIMO powinna wynosić właśnie połowę długości fali (ma to zapewnić brak wzajemnego oddziaływania elementów antenowych na siebie, a przez to możliwość traktowania kanałów propagacyjnych pomiędzy każdym elementem macierzy nadawczej a odbiornikiem jako niezależne);

- w tradycyjnym systemie wieloantenowym efektywna apertura antenowa jest dyskretna, w przypadku zaś macierzy HMIMOS można mówić o rozwiązaniach bliskich antenom o ciągłej aperturze (a jednocześnie pozwalających na uzyskanie wąskich wiązek propagacyjnych),
- w przypadku macierzy RMA (w wersji HMIMOS) liczba elementów składowych (*meta-atomów*) jest z założenia bardzo duża, w efekcie czego – przy bardzo niewielkiej odległości między elementami – efektywna apertura jest także duża.

Wspomniane powyżej różnice pomiędzy systemami tradycyjnymi a tymi wykorzystującymi komunikację holograficzną wskazują, że istotnie potrzeba prowadzenia intensywnych prac naukowych dotyczących różnych ich aspektów. Wystarczy tutaj wskazać np. na konieczność opracowania efektywnej i niezawodnej metody odwzorowywania informacji użytkownika w przestrzeń holograficzną, gdzie informacja przechowywana jest w postaci ekstremów pojawiających się w wyniku oddziaływania fali referencyjnej. W literaturze przedmiotu można znaleźć propozycje zastosowania zaawansowanych przetworników radiowo-optycznych pozwalających na wykorzystanie cech światła widzialnego w komunikacji holograficznej. W innych wariantach obraz holograficzny jest wytwarzany jednak w paśmie fal radiowych. Niezależnie jednak od tego, bardzo ważnym problemem staje się zaproponowanie wydajnych metod estymacji i korekcji kanału bezprzewodowego. Dodatkowo współcześnie dominującym rozwiązaniem w komunikacji holograficznej jest zapewnienie komunikacji do jednego ewentualnie niewielkiej liczby użytkowników; stąd istnieje konieczność dalszych badań w zakresie wielodostępu w systemach z komunikacją holograficzną. W istocie jednak potrzebne są także prace nad propozycją modelu propagacyjnego dla kanału bezprzewodowego, uwzględniające właściwości elektro-magnetyczne ośrodka względem propagującej się fali, a także nowe badania nad holograficzną teorią informacji, teorią kodowania czy synchronizacji. Warto tu podkreślić, że prace nad komunikacją holograficzną prowadzone są bardzo szeroko i niejednokrotnie proponowane rozwiązania wzbudzają wzmożoną dyskusję naukową. Jest to jednak obszar badawczy na tyle nowy a jednocześnie na tyle obiecujący, że autorzy wspomnianej już pracy [52] sugerują konieczność wypracowania nowej teorii transmisji danych, która będzie w sposób całościowy uwzględniała właściwości elektro-magnetyczne w systemie telekomunikacyjnym.

PODSUMOWANIE

Z perspektywy komunikacji bezprzewodowej przedstawione w artykule trzy rozwiązania technologiczne mogą być rozważane zarówno niezależnie, jak i łącznie. Meta-materiały, wytworzone z materiałów o ujemnym współczynniku załamania lub też z jednostek elementarnych (*meta-atomów*), mogą zostać wykorzystanie w celu wpływania na właściwości propagacyjne środowiska. Mogą też stanowić podstawę do wytworzenia rekonfigurowalnych powierzchni RMA, które z kolei mogą także zostać zbudowane w tradycyjny sposób z typowych obecnie elementów antenowych stosowanych w rozwiązaniach MIMO. Powierzchnie RMA dostarczają same w sobie ogromnych możliwości aplikacyjnych, a ich właściwości i możliwości zastosowania stały się w ostatnich jednym z najintensywniej rozważanych tematów naukowych na całym świecie w obszarze telekomunikacji bezprzewodowej. Wreszcie przedstawiona relatywnie niedawno propozycja całościowego wykorzystania właściwości wszystkich zjawisk pojawiających się podczas transmisji bezprzewodowej sygnałów wydaje się kreować zupełnie nowe obszary badawcze. Komunikacja holograficzna jako koncepcja transmisji danych jest dopiero w początkowej fazie rozwoju, a wiele aspektów związanych z przekazywaniem informacji w ten sposób wymaga intensywnych prac, wiele jest także rozbieżności oraz wątpliwością z nią związanych. Wydaje się jednak, że pojawiające się tutaj możliwości są na tyle obiecujące, że coraz częściej komunikację holograficzną wymienia się jako jedną z technik wykorzystywanych w przyszłych systemach bezprzewodowych. ●

PODZIĘKOWANIE

Praca została zrealizowana w ramach projektu nr 2021/43/B/ST7/01365 ufundowanego przez Narodowe Centrum Nauki (NCN).

LITERATURA

[1] M. Agiwal, A. Roy, N. Saxena, "Next Generation 5G Wireless Networks: A Comprehensive Survey", *IEEE Communications Surveys & Tutorials*, tom 18, nr 3, str. 1617-1655, trzeci kwartał 2016, doi: 10.1109/COMST.2016.2532458

[2] A. Osseiran et al., "Scenarios for 5G mobile and wireless communications: the vision of the METIS project", *IEEE Communications Magazine*, tom 52, nr 5, str. 26-35, May 2014, doi: 10.1109/MCOM.2014.6815890

[3] M. R. Palattella et al., "Internet of Things in the 5G Era: Enablers, Architecture, and Business Models", *IEEE Journal on Selected Areas in Communications*, tom 34, nr 3, str. 510-527, March 2016, doi: 10.1109/JSAC.2016.2525418

[4] J. Zhang, Y. Gao, X. Sun, W. Zhan, P. Liu, Z. Guo, "Synchronous Multi-Link Access in IEEE 802.11be: Modeling and Network Sum Rate Optimization", *ICC 2022 - IEEE International Conference on Communications*, Seul, Korea Południowa, 2022, str. 2309-2314, doi: 10.1109/ICC45855.2022.9838923

[5] Y. Liu, Y. Yu, Z. Du, L. Cuthbert, "Sequential State Q-learning Uplink Resource Allocation in Multi-AP 802.11be Network", *2022 IEEE 96th Vehicular Technology Conference* (VTC2022-Fall), Londyn, Wielka Brytania, 2022, str. 1-5, doi: 10.1109/VTC2022-Fall57202.2022.10013045

[6] M. Knitter, R. Kays, "Spatial Reuse Insights for IEEE 802.11ax and IEEE 802.11be Wireless LANs and Beyond", *2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio*

Communications (PIMRC), Kioto, Japonia, 2022, str. 919-925. doi: 10.1109/PIMRC54779.2022.9978080

[7] E. Khorov, I. Levitsky, I. F. Akyildiz, "Current Status, Directions of IEEE 802.11be, the Future Wi-Fi 7", *IEEE Access*, tom 8, str. 88664-88688, 2020.doi: 10.1109/ACCESS.2020.2993448

[8] C. Deng et al., "IEEE 802.11be Wi-Fi 7: New Challenges and Opportunities", *IEEE Communications Surveys & Tutorials*, tom 22, nr 4, str. 2136-2166, czwarty kwartał 2020. doi: 10.1109/COMST.2020.3012715

[9] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, M. Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications", *IEEE Communications Surveys & Tutorials*, tom 17, nr 4, str. 2347-2376, czwarty kwartał 2015, doi: 10.1109/COMST.2015.2444095

[10] L. Chettri, R. Bera, "A Comprehensive Survey on Internet of Things (IoT) Toward 5G Wireless Systems", *IEEE Internet of Things Journal*, tom 7, nr 1, str. 16-32, Jan. 2020. doi: 10.1109/JIOT.2019.2948888

[11] P. H. Pathak, X. Feng, P. Hu, P. Mohapatra, "Visible Light Communication, Networking, and Sensing: A Survey, Potential and Challenges", *IEEE Communications Surveys & Tutorials*, tom 17, nr 4, str. 2047-2077, czwarty kwartał 2015. doi: 10.1109/COMST.2015.2476474

[12] H. Haas, L. Yin, Y. Wang, C. Chen, "What is LiFi?", *Journal of Lightwave Technology*, tom 34, nr 6, str. 1533-1544, 15 March15, 2016. doi: 10.1109/JLT.2015.2510021

[13] W. Jiang, B. Han, M. A. Habibi, H. D. Schotten, "The Road Towards 6G: A Comprehensive Survey", *IEEE Open Journal of the Communications Society*, tom 2, str. 334-366, 2021, doi: 10.1109/OJCOMS.2021.3057679

[14] H. Al-Hraishawi, H. Chougrani, S. Kisseleff, E. Lagunas, S. Chatzinotas, "A Survey on Nongeostationary Satellite Systems: The Communication Perspective", IEEE Communications Surveys & Tutorials, tom 25, nr 1, str. 101-132, pierwszy kwartał 2023. doi: 10.1109/COMST.2022.3197695

[15] O. Kodheli et al., "Satellite Communications in the New Space Era: A Survey and Future Challenges", I*EEE Communications Surveys & Tutorials*, tom 23, nr 1, str. 70-109, pierwszy kwartał 2021. doi: 10.1109/COMST.2020.3028247

[16] W. Saad, M. Bennis, M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems", *IEEE Network,* tom 34, nr 3, str. 134-142, May/June 2020. doi: 10.1109/MNET.001.1900287

[17] P. Yang, Y. Xiao, M. Xiao, S. Li, "6G Wireless Communications: Vision and Potential Techniques", *IEEE Network*, tom 33, nr 4, str. 70-75, lipiec/sierpnień 2019. doi: 10.1109/MNET.2019.1800418

[18] W. Yang et al., "Semantic Communications for Future Internet: Fundamentals, Applications, and Challenges", I*EEE Communications Surveys & Tutorials*, tom 25, nr 1, str. 213-250, pierwszy kwartał 2023. doi: 10.1109/COMST.2022.3223224

[19] M. Polese, L. Bonati, S. D'Oro, S. Basagni, T. Melodia, "Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges", *IEEE Communications Surveys & Tutorials*, tom 25, nr 2, str. 1376-1411, drugi kwartał 2023, doi: 10.1109/COMST.2023.3239220

[20] https://openwifi.tip.build/openwifi-stack/openwifi-stack (dostęp 28 czerwca 2023)

[21] N. Hosseinidehaj, Z. Babar, R. Malaney, S. X. Ng, L. Hanzo, "Satellite-Based Continuous-Variable Quantum Communications: State-of-the-Art and a Predictive Outlook", IEEE Communications Surveys & Tutorials, tom 21, nr 1, str. 881-919, pierwszy kwartał 2019.doi: 10.1109/COMST.2018.2864557

[22] A. S. Cacciapuoti, M. Caleffi, R. Van Meter, L. Hanzo, "When Entanglement Meets Classical Communications: Quantum Teleportation for the Quantum Internet", *IEEE Transactions on Communications*, tom 68, nr 6, str. 3808-3833, czerwiec 2020, doi: 10.1109/TCOMM.2020.2978071

[23] P. Porambage, G. Gür, D. P. M. Osorio, M. Liyanage, A. Gurtov, M. Ylianttila, "The Roadmap to 6G Security and Privacy", I*EEE Open Journal of the Communications Society*, tom 2, str. 1094-1122, 2021, doi: 10.1109/OJCOMS.2021.3078081.

[24] W. Yuan et al., "New delay Doppler communication paradigm in 6G era: A survey of orthogonal time frequency space (OTFS)", *China Communications*, tom 20, nr 6, str. 1-25, czerwiec 2023, doi: 10.23919/JCC.fa.2022-0578.202306

[25] A. Liu et al., "A Survey on Fundamental Limits of Integrated Sensing and Communication", *IEEE Communications Surveys & Tutorials*, tom 24, nr 2, str. 994-1034, drugi kwartał 2022, doi: 10.1109/COMST.2022.3149272

[26] G. Liu, F. R. Yu, H. Ji, V. C. M. Leung, X. Li, "In-Band Full-Duplex Relaying: A Survey, Research Issues and Challenges", *IEEE Communications Surveys & Tutorials*, tom 17, nr 2, str. 500-524, drugi kwartał 2015, doi: 10.1109/COMST.2015.2394324

[27] A. O. Boryssenko, J. Liao, J. Zeng, S. Deng, V. M. Joyner, Z. R. Huang, "Radio-Optical Dual-Mode Communication Modules Integrated With Planar Antennas", *IEEE Transactions on Microwave Theory and Techniques*, tom 58, nr 2, str. 403-410, luty 2010, doi: 10.1109/TMTT.2009.2038441

[28] M. S. Saud, H. Chowdhury, M. Katz, "Heterogeneous Software-Defined Networks: Implementation of a Hybrid Radio-Optical Wireless Network", *2017 IEEE Wireless Communications and Networking Conference* (WCNC), San Francisco, CA, USA, 2017, str. 1-6., doi: 10.1109/WCNC.2017.7925873

[29] M. Kashef, M. Abdallah, N. Al-Dhahir, "Transmit Power Optimization for a Hybrid PLC/VLC/RF Communication System", I*EEE Transactions on Green Communications and Networking*, tom 2, nr 1, str. 234-245, marzec 2018, doi: 10.1109/TGCN.2017.2774104

[30] W. Liu, J. Ding, J. Zheng, X. Chen, C. -L. I, "Relay-Assisted Technology in Optical Wireless Communications: A Survey", *IEEE Access*, tom 8, str. 194384-194409, 2020, doi: 10.1109/ACCESS.2020.3031288

[31] Y. Liu et al., "Reconfigurable Intelligent Surfaces: Principles and Opportunities", *IEEE Communications Surveys & Tutorials*, tom 23, nr 3, str. 1546-1577, trzeci kwartał 2021., doi: 10.1109/COMST.2021.3077737

[32] Z. Ding et al., "A State-of-the-Art Survey on Reconfigurable Intelligent Surface-Assisted Non-Orthogonal Multiple Access Networks", *Proceedings of the IEEE*, tom 110, nr 9, str. 1358-1379, wrzesień 2022, doi: 10.1109/JPROC.2022.3174140

[33] A. C. Pogaku, D. -T. Do, B. M. Lee, N. D. Nguyen, "UAV-Assisted RIS for Future Wireless Communications: A Survey on Optimization and Performance Analysis", *IEEE Access*, tom 10, str. 16320-16336, 2022, doi: 10.1109/ACCESS.2022.3149054

[34] R. Deng et al., "Reconfigurable Holographic Surfaces for Future Wireless Communications", *IEEE Wireless Communications*, tom 28, nr 6, str. 126-131, grudzień 2021, doi: 10.1109/MWC.001.2100204

[35] R. Deng, Y. Zhang, H. Zhang, B. Di, H. Zhang, L. Song, "Reconfigurable Holographic Surface: A New Paradigm to Implement Holographic Radio", *IEEE Vehicular Technology Magazine*, tom 18, nr 1, str. 20-28, marzec 2023, doi: 10.1109/MVT.2022.3233157

[36] C. Huang et al., "Holographic MIMO Surfaces for 6G Wireless Networks: Opportunities, Challenges, and Trends", I*EEE Wireless Communications*, tom 27, nr 5, str. 118-125, październik 2020, doi: 10.1109/MWC.001.1900534

[37] Vladan Jevremovic, „Metamaterials and Reconfigurable Intelligent Surfaces", White paper, iBwave, dostęp online, (dostęp 28 czerwca 2023 r.)

[38] B. Ziętek, "Optoelektronika", 2011, Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika; Uzupełnienia do rozdziału IV.4.5, ISBN: 978-83-231-2746-8; dostęp online: http://fizyka.umk.pl/~bezet/uzupelnienia_opto.html (dostęp 28 czerwca 2023 r.)

[39] J. Franklin, J. Biddle, B. Balko „Double negative materials (DNM), Phenomena and Applications", IDA Document D-3887, 2009

[40] B. A. F. Esmail, S. Koziel, L. Golunski, H. B. A. Majid, R. K. Barik, "Overview of Metamaterials-Integrated Antennas for Beam Manipulation Applications: The Two Decades of Progress", *IEEE Access*, tom 10, str. 67096-67116, 2022, doi: 10.1109/ACCESS.2022.3185260.

[41] C. Liaskos, A. Tsioliaridou, A. Pitsillides, N. Kantartzis, A. Lalas, X. Dimitropoulos, S. Ioannidis, M. Kafesaki, C. Soukouli „Building Software Defined Materials with Source Nanonetworks", dostęp online: https://www.academia.edu/31120212/Building_Software_Defined_Materials_with_Nanonetworks (dostęp 28 czerwca 2023 r.)

[42] D. Manessis et al., "High Frequency Substrate Technologies for the Realisation of Software Programmable Metasurfaces on PCB Hardware Platforms with Integrated Controller Nodes", *2019 22nd European Microelectronics and Packaging Conference & Exhibition* (EMPC), Piza, Włochy, 2019, str. 1-7, doi: 10.23919/EMPC44848.2019.8951834.

[43] K. -K. Wong, K. -F. Tong, Z. Chu, Y. Zhang, "A Vision to Smart Radio Environment: Surface Wave Communication Superhighways", *IEEE Wireless Communications*, tom 28, nr 1, str. 112-119, luty 2021, doi: 10.1109/MWC.001.2000162.

[44] Ali, A.; Mitra, A.; Aïssa, B. Metamaterials and Metasurfaces: A Review from the Perspectives of Materials, Mechanisms and Advanced Metadevices. Nanomaterials, 2022, 12, 1027. https://doi.org/Source 10.3390/nano12061027

[45] Renzo, M.D., Debbah, M., Phan-Huy, D. et al. Smart radio environments empowered by reconfigurable AI meta-surfaces: an idea whose time has come. J Wireless Com Network 2019, 129 (2019). https://doi.org/10.1186/s13638-019-1438-9

[46] ETSI GR RIS 001 V1.1.1 (2023-04), "Reconfigurable Intelligent Surfaces (RIS); Use Cases, Deployment Scenarios and Requirements", https://www.etsi.org/committee/1966-ris, dostęp online: 28 czerwca 2023 r.

[47] ETSI GR RIS 003 V1.1.1 (2023-06), "Reconfigurable Intelligent Surfaces (RIS); Communication Models, Channel Models, Channel Estimation and Evaluation Methodology", https://www.etsi.org/committee/1966-ris, dostęp online: 28 czerwca 2023 r.

[48] H. Zhang, B. Di, "Intelligent Omni-Surfaces: Simultaneous Refraction and Reflection for Full-Dimensional Wireless Communications", *IEEE Communications Surveys & Tutorials*, tom 24, nr 4, str. 1997-2028, czwarty kwartał 2022, doi: 10.1109/COMST.2022.3202813.

[49] R. Liu, J. Dou, P. Li, J. Wu, Y. Cui, "Simulation and Field Trial Results of Reconfigurable Intelligent Surfaces in 5G Networks", *IEEE Access*, tom 10, str. 122786-122795, 2022, doi: 10.1109/ACCESS.2022.3223447.

[50] Wankai Tang, Xiang Li, Jun Yan Dai, Shi Jin, Yong Zeng, Qiang Cheng, Tie Jun Cui. Wireless Communications with Programmable Metasurface: Transceiver Design and Experimental Results. China Communications. 2019, 16(5): 46-61

[51] D. Dardari, N. Decarli, "Holographic Communication Using Intelligent Surfaces", *IEEE Communications Magazine*, tom 59, nr 6, str. 35-41, czerwiec 2021, doi: 10.1109/MCOM.001.2001156.

[52] Tierui Gong, Panagiotis Gavriilidis, Ran Ji, Chongwen Huang, George C. Alexandropoulos, Li Wei, Zhaoyang Zhang, Mérouane Debbah, H. Vincent Poor, Chau Yuen, "Holographic MIMO Communications: Theoretical Foundations, Enabling Technologies, and Future Directions", https://arxiv.org/abs/2212.01257, https://doi.org/10.48550/arXiv.2212.01257

[53] C. Huang et al., "Holographic MIMO Surfaces for 6G Wireless Networks: Opportunities, Challenges, and Trends", *IEEE Wireless Communications*, doi: 10.1109/MWC.001.1900534

[54] Rajatheva, N., Atzeni, I., Björnson, E., Bourdoux, A., Buzzi, S., Doré, J.-B., Erkucuk, S., Fuentes, M., Guan, K., Hu, Y., Huang, X. , Hulkkonen, J., Jornet, J. M., Katz, M., Nilsson, R., Panayirci, E., Rabie, K., Rajapaksha, N., Salehi, M., … Xu, W. (2020). White Paper on Broadband Connectivity in 6G [White paper]. (6G Research Visions, Nr 10). University of Oulu. http://urn.fi/ urn:isbn:9789526226798

[55] R. Deng et al., "Reconfigurable Holographic Surfaces for Future Wireless Communications", *IEEE Wireless Communications*, tom 28, nr 6, str. 126-131, grudzień 2021, doi: 10.1109/MWC.001.2100204.

[56] Z. Yang, Y. Hu, Z. Zhang, W. Xu, C. Zhong, K. -K. Wong, "Reconfigurable Intelligent Surface Based Orbital Angular Momentum: Architecture, Opportunities, and Challenges", I*EEE Wireless Communications*, tom 28, nr 6, str. 132-137, grudzień 2021, doi: 10.1109/MWC.001.2100223.

DOI: 10.15199/59.2023.4.5

Wyzwania społeczne, ekonomiczne i techniczne wdrożenia transportu autonomicznego

Social, economic and technical challenges of autonomous transport implementation

STRESZCZENIE: W tekście omawiane są wyzwania społeczne, ekonomiczne i techniczne związane z wdrażaniem transportu autonomicznego. Transport autonomiczny obejmuje różne rodzaje pojazdów, takie jak drogowe, szynowe, pływające i latające. Istnieją zarówno czynniki przemawiające za robotyzacją transportu, takie jak poprawa bezpieczeństwa i ograniczenie kosztów, jak i przeciw, takie jak błędy oprogramowania i nadmierne wykorzystanie danych użytkowników. Przełom w zakresie implementacji transportu autonomicznego przewiduje się w ciągu najbliższych sześciu lat, zwłaszcza w latach 2027–2030. W przypadku pojazdów szynowych, takich jak pociągi, tramwaje i metro, nie ma większych problemów prawnych ani technologicznych, a ich wdrożenie pozwala na optymalizację czasu jazdy, zwiększenie prędkości systemu i skrócenie czasu postoju na stacjach. Natomiast w przypadku samochodów, statków i dronów istnieją jeszcze pewne wyzwania, takie jak bezpieczeństwo i regulacje prawne. Warto również zauważyć, że rozwój infrastruktury teleinformatycznej, w tym transmisji danych w technologii 5G, jest kluczowy dla skutecznej implementacji pojazdów autonomicznych.
SŁOWA KLUCZOWE: drony, roboty, statki, transport autonomiczny, wyzwania

ABSTRACT: The text discusses the social, economic, and technical challenges of implementing autonomous transport. Autonomous transport includes different types of vehicles, such as road, rail, floating and flying. There are both factors in favour of the robotisation of transport, such as improved safety and cost reduction, and against, such as software bugs and overuse of user data. A breakthrough in the implementation of autonomous transport is predicted in the next six years, especially between 2027 and 2030. For rail vehicles such as trains, trams and metros, there are no major legal or technological issues, and their implementation allows for optimised driving times, increased system speeds and reduced station stops. In contrast, there are still some challenges with cars, ships, and drones, such as safety and regulation. It is also worth noting that the development of ICT infrastructure, including 5G data transmission, is crucial for the successful implementation of autonomous vehicles.
KEYWORDS: drones, robots, ships, autonomous transport, challenges

red. SŁAWOMIR KOSIELIŃSKI
Fundacja Instytut Mikromakro, Collegium Civitas, Warszawa,
kosiel@mikromakro.pl

Transport autonomiczny obejmuje pojazdy drogowe, szynowe (kolej, tramwaje, metro), pływające (promy, statki, jachty), samoloty (załogowe statki powietrzne z funkcją autopilota) oraz drony (bezzałogowe statki powietrzne). Każdy z tych rodzajów środków transportów ma własną kategoryzację stopni autonomii. Mimo to wyzwania społeczne, ekonomiczne i techniczne, a nade wszystko prawne, są do siebie dość zbliżone.

Za robotyzacją transportu przemawiają takie czynniki jak:

- poprawa bezpieczeństwa;
- rozwiązanie problemu deficytu kierowców zawodowych, maszynistów, marynarzy i pilotów;
- ekologiczne pojazdy;
- szybka dostawa towarów;
- ograniczenie kosztów;
- odciążenie pracowników w ich codziennej pracy.

Przeciwko są obawy, że sprzęt może zawieść w kluczowych momentach np. z powodów błędów oprogramowania czy niedostatecznej jakości transmisji danych spowodowanej lukami w infrastrukturze dostępowej. Nie można przełożyć na kod maszynowy doświadczenia i intuicji kolejarzy, marynarzy, pilotów i kierowców. Do tego dochodzi niepokój o sposób wykorzystania nagromadzonych znacznej ilości danych o użytkownikach i ich zachowaniu. Rośnie pokusa nadużyć, naruszeń prywatności i zależności od firm technologicznych. Ponadto robotyzacja środków transportu ma powodować utratę kontroli i interakcji międzyludzkich.

Bezduszny algorytm zdecyduje za nas, kogo wyprzedzić albo kiedy zwolnić. Przypuszcza się, że wpłynie to na pogorszenie jakości doświadczenia podróżowania.

Od strony ekonomicznej padają dodatkowe argumenty przeciw robotyzacji transportu. Koszty wdrożenia transportu autonomicznego mogą się okazać zbyt wielkie w stosunku do korzyści. Nie tylko same pojazdy autonomiczne są drogie, ale również wymagają odpowiedniej infrastruktury teleinformatycznej i systemów wsparcia. Mimo to kolejne raporty rynkowe szacują gwałtowny wzrost wartości globalnego rynku pojazdów autonomicznych z poziomu 10,1 mld USD w 2021 roku do 325,9 mld USD w 2030 roku [10].

Ta zależność od technologii rodzi obawy i lęki społeczne jakże przypominające luddyzm – ruch wśród robotników i rzemieślników angielskich na przełomie XVIII i XIX w., którzy niszczyli maszyny, ponieważ widzieli w nich główną przyczynę niskich płac i zagrożenia bezrobociem. W rezultacie wzrost techno-pesymizmu [8] skutkuje decyzjami politycznymi i administracyjnymi utrudniającymi lub wręcz uniemożliwiającymi rozwój transportu autonomicznego. Przykładowo Agencja Unii Europejskiej ds. Bezpieczeństwa Lotniczego (EASA), zaostrzając przepisy regulujące loty dronów, próbuje powstrzymać rozwój sektora małych i średnich przedsiębiorstw korzystających z latających robotów, które stają się konkurencyjne wobec lotnictwa załogowego. Jednocześnie preferuje tzw. bezzałogowe taksówki powietrzne projektowane przez wielkie koncerny lotnicze i motoryzacyjne lub wspierane przez nich startupy, zakładając, że te maszyny przejdą kosztowny proces certyfikacji w jej laboratoriach [5].

Natomiast prawo o ruchu drogowym wciąż nie dopuszcza ruchu pojazdów autonomicznych po drogach publicznych, blokując (na razie) rozwój tej branży. I znowu warto odwołać się do historii. Ustawa o czerwonej fladze – potoczna nazwa ustawy *Locomotives on Highways Act* z 1865 roku – wprowadziła w Wielkiej Brytanii zakaz poruszania się omnibusów parowych bez osoby biegnącej 55 metrów przed pojazdem i ostrzegającej innych użytkowników dróg. Musiało minąć kilkadziesiąt lat jak rozwój techniki motoryzacyjnej odesłał ten akt prawny do lamusa.

I właśnie poziom gotowości technologicznej (ang. *Technology Readiness Level,* TRL) transportu autonomicznego, który gremialnie oscyluje między poziomem 7 (testy w warunkach operacyjnych) a poziomem 8 (demonstracja ostatecznej wersji technologii) jest tutaj najistotniejszy. Gdy pojawią się maszyny na poziomie 9 – technologii gotowej do wdrożenia, zaniknie zdecydowana większość barier.

POJAZDY DROGOWE

Zielony bus skonstruowany w gliwickiej firmie Blees ma 2,8 metra wysokości, 2 metry szerokości i 5 metrów długości. Jest w nim miejsce, żeby przewozić do 15 osób na 7 miejscach siedzących i 8 stojących. Całość została dostosowana do osób niepełnosprawnych. W stosunku do innych pojazdów wyróżnia go, że na pokładzie nie ma kierowcy. To bus autonomiczny, jedynie nadzorowany zdalnie. Operator wkraczałby do akcji, gdyby system nie poradził sobie z sytuacją drogową.

W ocenie firmy bus ułatwi transport na obrzeżach miast, w których pełno jest tzw. białych plam transportowych. Wysokie koszty operacyjne dla przewoźników i pogłębiający się brak kierowców zdecydowanie utrudniają poprawę sytuacji. Autonomiczny bus ma uzupełniać sieć transportu publicznego i w żadnym wypadku nie zabierać pracy kierowcom. Ma jeździć wszędzie tam, gdzie jest potrzeba, bez sztywnych rozkładów jazdy. Mieszkaniec lub gość nowego osiedla na obrzeżu miast albo stref czystego transportu zamówi przejazd tym busem w specjalnej aplikacji. Wybierze godzinę i uda się na miejsce odbioru. To istotnie poprawi jakość życia zwłaszcza osób starszych.

Pilotaże transportu na żądanie trwają od dwóch lat w Sosnowcu, Międzyzdrojach i Skoczowie. Wkrótce dojdą kolejne. Ograniczenia prawne zmuszają wybór tras w miejscach zamkniętych, nieposiadających statusu drogi publicznej jak np. cmentarze czy parki. Powodem jest brak homologacji pierwszej wersji busa. Dopiero ok. roku 2026/2027 powstanie druga wersja i wtedy przejdzie cały proces homologacyjny. W rezultacie bus trafi do masowej produkcji i sprzedaży.

Poziomy autonomiczności

Punktem odniesienia, w jakim miejscu znajduje się pojazd firmy Blees, jest ocena poziomu autonomiczności w samochodach opracowana przez Amerykańską Narodową Administrację Bezpieczeństwa Ruchu Drogowego (NHTSA) [9]. Pozwala to na ustandaryzowanie możliwości i funkcjonalności w poszczególnych pojazdach w przejrzysty sposób. Poniższy podział na sześć poziomów od 0 (w pełni manualny) do 5 (w pełni autonomiczny) przyjęty został również przez amerykańskie Stowarzyszenie Inżynierów Motoryzacji (SAE) i inne stowarzyszenia branżowe. W konsekwencji został uznany za globalny standard.

- *Poziom 0.* Większość obecnie poruszających się pojazdów jest na 0 poziomie autonomiczności. To prosta, sterowana ręcznie technika. Kierowca nie jest wspomagany przez żadne zaawansowane technologie, choć pomagają mu pewne systemy, prf.. hamowania awaryjnego.
- *Poziom 1.* Najniższy poziom autonomizacji. Pojazd jest wyposażony w pojedynczy zautomatyzowany system wspomagający kierowcę, na przykład układ kierowniczy lub tempomat.
- *Poziom 2.* Pojazdy są wyposażone w zaawansowane systemy, które mogą kontrolować zarówno sterowanie, jak i przyspieszanie oraz zwalnianie. Na tym poziomie samochody jeszcze nie są w pełni autonomiczne, ponieważ jest wymagana obecność człowieka. Kierowca może przejąć kontrolę nad samochodem w dowolnym momencie.
- *Poziom 3.* Pojazdy monitorują otoczenie i innych uczestników ruchu. Mogą samodzielnie podejmować pewne decyzje, np. przyspieszanie obok wolno jadącego się samochodu. W dalszym ciągu jednak wymagają kontroli ze strony człowiek – kierowca musi zachować czujność i być gotowy do przejęcia kierowania, jeśli system nie jest w stanie wykonać zadania.

- *Poziom 4.* Udział człowieka nie jest potrzebny, ale w kabinie w dalszym ciągu istnieje miejsce dla kierowcy, który może poprowadzić samochód ręcznie. Ciężarówki tego typu mogą poruszać się samodzielnie, ale – póki nie zostanie stworzona odpowiednia infrastruktura i przepis - mogą jeździć tylko na ograniczonym obszarze.
- *Poziom 5.* Pełna autonomiczność. Nie jest wymagany ludzki czynnik ani kontrola zdalna. Pojazd nie jest w żaden sposób ograniczony (np. geofencingiem, w ramach którego tworzy się pewne strefy, przekroczenie których jest niemożliwe lub ściśle kontrolowane). Jakkolwiek ciężarówki o poziomie autonomiczności 1–4 są już w użytku, to pojazdy z poziomu 5 nie są dopuszczone do ruchu nigdzie na świecie.

W rezultacie samochody osobowe marki Tesla, które w oczach opinii publicznej uchodzą za autonomiczne, są pomiędzy poziomem 2 a poziomem 3. Ale bus gliwickiej spółki Blees jest dzisiaj na poziomie 4 z perspektywą szybkiego rozwoju w stronę pełnej autonomiczności na poziomie 5. Jednak w dyskursie publicznym pojawiają się głosy, że wcale nie jest pewne, iż w ogóle osiągniemy ten poziom [4]. Być może zawsze człowiek będzie musiał kontrolować przejazd, aby w razie czego przejąć sterowanie. Okazuje się, że sztuczna inteligencja (ang. *AI*) nie daje sobie rady ze wszystkimi uwarunkowaniami ruchu drogowego, czytaj – nie potrafimy napisać w pełni bezpiecznego i efektywnego programu komputerowego, którym *de facto* jest AI.

Technologie poziomu 3 są gotowe od 2017 r., lecz producenci samochodów nie udostępniali tych autonomicznych funkcji ze względu na przepisy, które nie definiowały, jak ma sprawnie działać taki auto-robot i nie określały odpowiedzialności w razie wypadku. Czy winien jest kierowca, który w porę nie przejął kontroli nad samochodem, czy też programista, który źle postawił przecinek w kodzie oprogramowania sterującego hamowaniem w trybie alarmowym? I jeszcze jedno: jeśli dojdzie do wypadku, w którym ani chybi musi ktoś zginąć, by inni ocaleli, czyje życie jest ważniejsze: kierowcy czy babci, goniącej wnuczka, beztrosko wbiegającego na jezdnię na czerwonym świetle? Zdanierof.of. Marcina Ślęzaka, dyrektora Instytutu Transportu Samochodowego, odpowiedź na te etyczne dylematy zależy od kręgu kulturowego. Inaczej odpowiadają Amerykanie, inaczej Europejczycy czy też mieszkańcy Chin. Zatem każdy samochód autonomiczny musi mieć jeszcze wgrany program etyczny zgodnie z obyczajami danego rynku [4].

Czynniki ekonomiczne

O ile rozwój autonomicznych pojazdów osobowych może się okazać ślepą uliczką, ponieważ ani nie zmniejszy ruchu drogowego ani nie ograniczy śladu węglowego, o tyle bardzo obiecująco wygląda wdrożenie autonomicznych ciężarówek i wspomnianych wyżej busów.

Otóż, jednym z największych problemów nękających obecnie branżę transportową jest ogromny niedobór siły roboczej. W 2021 r. American Trucking Associations oszacowało, że branża borykała się z historycznym niedoborem 80 tys. kierowców, przy czym oczekiwano, że niedobór może przekroczyć 160 tys. kierowców do 2030 r., biorąc pod uwagę obecne dane demograficzne kierowców i oczekiwany wzrost wolumenu przewozów [1]. Sytuacja jest bardzo podobna w Unii Europejskiej, gdzie w 2021 roku brakowało 400 tys. kierowców samochodów ciężarowych.

Notabene, według danych Eurostatu (Urzędu Statystycznego Unii Europejskiej) przewozy transgraniczne w ciągu każdego dnia roboczego w Europie realizuje ok. 600 tys. „ekwiwalentów pełnych etatów" pojazdów ciężarowych (z łącznej liczby ok. 2,6 mln samochodów ciężarowych działających w Europie, oraz z zarejestrowanych łącznie 7,5 mln ciężarówek i ciągników o DMC > 3,5 tony, czyli o dopuszczalnej masie całkowitej, którą się definiuje zgodnie z prawem o ruchu drogowym w Polsce jako największą, określoną właściwymi warunkami technicznymi masę pojazdu, obciążonego osobami i ładunkiem, dopuszczonego do poruszania się po drodze).

Obsługuje je ok. 742 tys. pełnoetatowych kierowców międzynarodowych, którzy, jeśli mieli odpowiednie doświadczenie i szczęście trafili do nielicznej grupy firm lepiej honorujących swój personel. Ale wzrost zapotrzebowania na transport drogowy i systemowy brak kierowców w Europejskim Obszarze Gospodarczym oraz Wielkiej Brytanii powoduje, że przedsiębiorcy zaczynają coraz częściej podnosić stawki godzinowe i inwestować w specjalistyczne szkolenia pracowników, byle chcieli u nich jeździć.

Wbrew pozorom nie powoduje to gwałtownych ruchów kadrowych. Przepływy między firmami są niewielkie, kierowcy cenią sobie stabilność pracy i solidność swoich pracodawców. Na zmiany decydują się młodsi kierowcy, którzy chcą się szybko dorobić, by stać ich było na zakup własnego ciągnika osiowego (czyli pojazd ciężarowy z naczepą) i móc rozpocząć samodzielna pracę.

Sam polski rynek transportu drogowego liczy ok. 33 tys. firm. W ocenie Polskiego Instytutu Transportu Drogowego (PITD) chociaż pracuje w nich ponad 650 tys. kierowców i liczba ta ciągle rośnie, popyt pozostaje dalej niezaspokojony. Siedem lat temu polski oddział PwC przygotował raport „Rynek pracy kierowców w Polsce" [11]. Już wówczas szacowano, że na rynku pracy brakuje ok. 100 tys. zawodowych kierowców. Obecnie brakuje w Polsce około 140 tys. kierowców zawodowych, zaś do 2025 r. ta luka powiększy się do 300 tys. wg. szacunków Międzynarodowej Unii Transportu (IRU).

- Pandemia koronawirusa obnażyła bezlitośnie braki kadrowe – chorego pracownika nie było kim zastąpić, ponieważ ew. zastępcy nie mieli potrzebnych uprawnień i licencji. Do dzisiaj aż 43% polskich firm transportowych ma problem ze znalezieniem wykwalifikowanego pracownika. Kierowcy są obecnie głównym wąskim gardłem dla przepustowości całego europejskiego systemu transportowego. Zdaniem prof. Witolda Orłowskiego, głównego doradcy ekonomicznego PwC, *„wykorzystanie potencjału rozwojowego Polski wymaga kształcenia nie tylko w „zawodach przyszłości", ale i w zakresie bardziej powszechnego zawodu kierowcy".*[11].

Brak wystarczającej liczby kierowców oraz konieczność wymiany silników spalinowych na elektryczne lub wodorowe są silnym bodźcem do poszukiwania innowacji techno-

logicznych i społecznych, w tym do inwestycji w transport autonomiczny. Takimi czynnikami są również warunki pracy kierowców ujęte w unijnym Pakiecie Mobilności (2020 r.), które wpływają w istotny sposób na koszty świadczenia usług transportowych.

Celem było unormowanie transportu drogowego na terenie Unii Europejskiej poprzez zmianę obowiązujących przepisów w kluczowych kwestiach:

- czas jazdy i odpoczynku oraz tachografy;
- delegowanie kierowców;
- dostęp do rynku.

Już w sierpniu 2020 r. wszedł w życie przepis, który nakazuje przedsiębiorcy zapewnić kierowcy powrót do centrum operacyjnego (bazy) lub miejsca zamieszkania kierowcy co 4 tygodnie:

- powrót na co najmniej 45 godzin;
- w przypadku wykorzystania w trasie dwóch odpoczynków tygodniowych skróconych – powrót przed kolejnym odpoczynkiem tygodniowym regularnym (co 3 tygodnie).

Nie wolno przy tym odpoczywać kierowcy powyżej 45 godzin w kabinie pojazdu. Musi być to miejsce przyjazne dla wszystkich płci z odpowiednim zapleczem sanitarnym oraz sypialnym. Generalnie skończyły się jazdy po całej Europie bez solidnego odpoczynku. Każdy przejazd i postój są monitorowany przez coraz bardziej zaawansowane tachografy – urządzenia elektroniczne do kontroli pracy kierowcy. Natomiast od 2 lutego 2022 r. kierowcy muszą otrzymywać pełne wynagrodzenie minimalne adekwatne do kraju, w którym wykonują usługę. Wyłączone z tego są przejazdy tranzytowe, zaś także transport bilateralny z i do Polski. Dotyczy to także przejazdów kabotażowych, czyli jeśli kierowca polskiej firmy przewozi towar na terenie innego kraju Unii lub cross trade, czyli z jednego państwa unijnego do drugiego, innego niż siedziba przewoźnika.

Oto przykłady. Jeśli kierowca wykonuje przewóz towarów we Włoszech z pkt. A do pkt. B (przewóz kabotażowy), to podlega on nowym zasadom delegowania i jego wynagrodzenie musi być większe lub równe wynagrodzeniu kierowcy włoskiego. Ale jeśli zajmuje się przewozem towaru z Włoch do Francji (przewóz *cross-trade*), to wówczas podlega on nowym zasadom delegowania i jego wynagrodzenie musi być większe lub równe wynagrodzeniu kierowcy włoskiego za część godzin przepracowanych we Włoszech oraz wynagrodzeniu kierowcy francuskiego za część godzin przepracowanych we Francji. Jeśli zaś zadaniem kierowcy jest przewóz towarów z Polski do Francji (przewóz dwustronny), przejeżdżając przez Niemcy (tranzyt), to jego wynagrodzenie na całej długości trasy wynika jedynie z polskiego ustawodawstwa. Za brak spełnienia dotyczących obowiązków delegowania, nie grożą jedynie kary finansowe, ale także utrata dobrej reputacji przewoźnika, a co za tym idzie nawet zawieszenie licencji.

Nowe przepisy wymusiły na przewoźnikach ponowne skalkulowanie cen swoich usług. Szacuje się, że w zależności od typu wykonywanych przewozów wzrost kosztów wyniósł od 20% do nawet 100%. Przykładowo, jeżeli w 2021 roku kierowca zarabiał 7 tys. zł „na rękę", a podstawa wynosiła 3600 zł brutto i reszta była dopełniona nieopodatkowanymi ryczałtami i dietami, to łączny koszt pracodawcy wynosił 8700 zł brutto. Przy założeniu, że poza granicami kraju wykonuje więcej przewozów podlegających przepisom pakietu mobilności, to od lutego 2022 roku ten koszt, mógł wzrosnąć nawet do 14 tys. zł brutto.

To kolejny sygnał, że przedsiębiorcy powinni nastawić się na przyspieszenie cyfryzacji transportu i transport autonomiczny. Inaczej sektor transportu może się znaleźć na skraju załamania. Przykładowo według danych Głównego Urzędu Statystycznego z 2018 roku, puste przebiegi stanowią 34% wszystkich przejazdów w Polsce i 12%. w transporcie międzynarodowym. Połowa realizowanych przewozów nie wykorzystuje całkowitej pojemności ciężarówek. Założeniem zrobotyzowanych pojazdów jest zatem szeroko zakrojona optymalizacja - redukcja przestojów, zarządzanie procesami logistycznymi, likwidacja dublujących się przejazdów.

Kadry na potrzeby transportu drogowego w technologii platooningu i autopilotażu

Koncerny motoryzacyjne i firmy robotyczne pokładają nadzieję w konstrukcji pojazdów autonomicznych i połączonych. O ile na w pełni zautomatyzowany transport poczekamy zapewne do końca tej dekady, o tyle truck platooning w wersji komercyjnej pojawi się rzeczywiście za dwa-trzy lata. To synchroniczna jazda kolumny ciężarówek, które podążają za sobą ok. 10 metrów, korzystając z technologii łączności i zautomatyzowanych systemów wspomagających jazdę.

Przynieść ma to wiele korzyści, między innymi zmniejszenie zużycia paliwa, a w konsekwencji emisji CO2. Zwiększyć ma się także bezpieczeństwo na drogach. Gdy pierwszy samochód ciężarowy zahamuje, niemal w tej samej chwili robią to pojazdy za nim – reakcja maszyn jest pięciokrotnie szybsza niż człowieka.

Siłą rzeczy warto przygotować już dzisiaj kadry, które będą zawiadywać takim transportem. Takie osoby będą nadzorować zdalnie na początku platooning (transport połączony), z czasem zarządzać ruchem ciężarówek autonomicznych. Potrzebują oni umiejętności oceny warunków ruchu, nie tylko pod kątem zachowania się innych kierowców, ale także pod kątem możliwych reakcji algorytmów sterujących samochodem autonomicznym. Niezbędna jest także gotowość do szybkiej reakcji w razie wystąpienia niebezpieczeństwa błędnego działania funkcji autonomicznych. W tym celu niezbędna jest wiedza, zarówno teoretyczna, jak i praktyczna, obejmująca zasady działania funkcji jazdy automatycznej, ich możliwości i ograniczeń. Konieczna jest także świadomość zagrożeń, jakie może spowodować niewłaściwe wykorzystanie funkcji autonomicznych lub zbyt duże zaufanie do możliwości tych systemów [3].

Ponadto umiejętność zdalnego nadzoru nad samochodami ciężarowymi o nawet częściowej autonomizacji jazdy otwiera możliwość pracy dla osób z niepełnosprawnościami, którzy albo byli kierowcami albo chcieliby dalej pracować w zbliżonej specjalności.

Akceptacja społeczna

Warunkiem sine qua non rozwoju pojazdów autonomicznych jest uzyskanie dlań większej akceptacji społecznej i zaufania. Dzisiaj jest ono dalekie od oczekiwań techno-optymistów [8] uznających tą technologię za przyszłość mobilności. Wynika to z badania J.D. Power *2022 US Mobility Confidence Index* (MCI) [2]. Według niego zrozumienie przez amerykańskich konsumentów pojazdów zautomatyzowanych jest nie tylko niskie, ale w rzeczywistości spadło od 2021 r.

Chociaż temat rozwoju pojazdów autonomicznych jest od dawna obecny w debacie publicznej, badanie wykazało, że 65% konsumentów błędnie określa, czymże są w pełni zautomatyzowane, samojezdne pojazdy. Co więcej, 56% respondentów niewłaściwie sklasyfikowało dostępne obecnie technologie wspomagające kierowcę jako w pełni zautomatyzowane technologie samojezdne, co świadczy o znacznej luce w wiedzy o pojazdach autonomicznych i ogólnym braku zrozumienia w przygotowaniu do wyższych poziomów automatyzacji.

Dodatkowym wyzwaniem jest to, że wielu konsumentów nie potrafi odróżnić wspomagania jazdy, wspomagania kierowcy i półautonomii, z których każda różni się definicją i zastosowaniem. Jednak dla wielu terminy te łączą się ze sobą i powodują zamieszanie w szerszym kontekście automatyzacji, zgodnie z definicjami jazdy autonomicznej NHTSA. Tym samym nadal producenci samochodów muszą uczyć swoich klientów o możliwościach autonomicznych pojazdów.

Należy tutaj wyjaśnić, że o ile z perspektywy takich firm jak Blees liczy się transport publiczny jako odbiorca ich pojazdów, o tyle badanie MCI skupia się na użytkowniku indywidualnym (4000 właścicieli pojazdów w USA powyżej 19 roku życia). Ci wymagają dodatkowych szkoleń i akcji popularyzacyjnych. Inaczej wciąż będą wykazywać w kolejnych badaniach niski poziom komfortu i pewności podczas jazdy w całkowicie zautomatyzowanym pojeździe.

Jednakże w skali świata wygląda to już zdecydowanie korzystniej. Dwa lata przed badaniem MCI ukazał się raport KPMG „Autonomous Vehicles Readiness Index 2020" [6]. To badanie gotowości krajów na wdrożenie autonomicznych pojazdów. W raporcie przedstawione są wyniki badań dotyczące czterech filarów: polityki i przepisów, technologii i innowacji, infrastruktury oraz akceptacji konsumentów z 30 krajów świata.

Krajem, który uzyskał najwyższy wynik i jest najbardziej gotowy na wdrożenie transportu autonomicznego, jest Holandia. Po nim lokują się Singapur, Stany Zjednoczone, Szwecja i Wielka Brytania. Natomiast ostatnie miejsca zajmują Brazylia, Rosja, Meksyk i Indie. To pochodna stanu infrastruktury teleinformatycznej. Im lepsza i bardziej wydajna sieć transmisji danych, tym większa szansa na skuteczną implementację pojazdów autonomicznych.

POJAZDY SZYNOWE

O ile autonomiczne pojazdy drogowe wzbudzają nadmierne emocje, to roboto-pociągi, tramwaje i metro cieszą się pełną akceptacją społeczną. Nie ma też większych problemów prawnych ani technologicznych. To tylko kwestia rachunku ekonomicznego. Ten zaś jest bardzo korzystny, zwłaszcza w przypadku kolejek miejskich i pociągów dużej prędkości. Wdrożenie szynowych systemów autonomicznych pozwala operatorom na optymalizację czasu jazdy pociągów, zwiększenie średniej prędkości systemu, skrócenie przejazdów do 75 sekund oraz skrócenie czasu postoju na stacjach (w optymalnych warunkach) do 15 sekund. Mimo to tylko 6% długości linii metra na świecie jest w pełni autonomiczna.

Dzięki wyeliminowaniu czynnika ludzkiego z równania prowadzenia, operatorzy zyskują elastyczność i mogą lepiej wykorzystywać zasoby. Takie systemy dają bardziej zindywidualizowany zakres usług, zmniejszając nadmiar mocy produkcyjnych poza godzinami szczytu i umożliwiając operatorom wprowadzanie pociągów w odpowiedzi na nagłe wzrosty popytu, np. w przypadku dużych wydarzeń sportowych, kulturalnych czy religijnych.

Niewątpliwie ważnym czynnikiem ekonomicznym jest brak chętnych do pracy na kolei. W samych tylko Niemczech brakuje ok. 1500 maszynistów. Automatyzacja wydaje się optymalnym rozwiązaniem.

Poziomy autonomiczności

Według Międzynarodowego Stowarzyszenia Transportu Publicznego (UITP) istnieje pięć stopni automatyzacji pociągów (GoA) z podziałem na sposób eksploatacji i opisem:

- GoA 0 → w zasięgu wzroku → brak automatyzacji np. w tramwaju, wobec czego motorniczy musi samodzielnie zajmować się całym procesem jazdy, zatrzymywania się i otwierania drzwi.
- GoA 1 → ręczne sterowanie → maszynista kontroluje uruchamianie i zatrzymywanie, działanie drzwi i przejmowanie sterowania w nagłych przypadkach.
- GoA 2 → półautomat (STO) → uruchamianie i zatrzymywanie są zautomatyzowane, ale maszynista obsługuje drzwi, w razie potrzeby prowadzi pociąg i zajmuje się sytuacjami awaryjnymi. Wiele systemów automatycznej obsługi pociągów (ang. ATO) to GoA 2.
- GoA 3 → bez maszynisty (DTO) → ruszanie i zatrzymywanie są zautomatyzowane, ale steward obsługuje drzwi i prowadzi pociąg w sytuacjach awaryjnych.
- GoA 4 → eksploatacja pociągów bezobsługowych (ang. UTO: *Unattended Train Operation*) → uruchamianie i zatrzymywanie, obsługa drzwi jest w pełni zautomatyzowana bez personelu w pociągu. Zaleca się, aby stacje miały zamontowane drzwi ekranowe peronu.

W systemach metra automatyzacja odnosi się do procesu, w którym odpowiedzialność za zarządzanie eksploatacją pociągów jest przenoszona z maszynisty na system sterowania pociągiem (metroautomation.org/).

Postęp techniczny sprawił, że systemy sterowania pociągami są w stanie nadzorować, obsługiwać i kontrolować cały proces eksploatacyjny. Kluczowe elementy to:

- Automatyczna ochrona pociągu (ATP) to system i wszystkie urządzenia odpowiedzialne za podstawowe bezpieczeństwo; unika kolizji, przekraczania czerwonego sygnału i przekraczania ograniczeń prędkości dzięki automatycznemu

hamowaniu. Linia wyposażona w ATP odpowiada (przynajmniej) GoA1.

- Automatyczna obsługa pociągów (ATO) zapewnia częściowe lub całkowite automatyczne pilotowanie pociągu i funkcje bez maszynisty. System ATO wykonuje wszystkie funkcje kierowcy, z wyjątkiem zamykania drzwi. Maszynista musi tylko zamknąć drzwi, a jeśli droga będzie wolna, pociąg automatycznie przejedzie na następną stację. Odpowiada to GoA2. Wiele nowszych systemów jest całkowicie sterowanych komputerowo; większość systemów nadal decyduje się na utrzymanie maszynisty lub jakiegoś rodzaju stewardesy, aby złagodzić ryzyko związane z awariami lub sytuacjami awaryjnymi. Odpowiada to GoA3.
- Automatyczne sterowanie pociągiem (ATC) wykonuje automatycznie normalne operacje sygnalizatora, takie jak wyznaczanie trasy i regulacja pociągu. Systemy ATO i ATC współpracują ze sobą, aby utrzymać pociąg w określonej tolerancji rozkładu jazdy. Połączony system nieznacznie dostosuje parametry operacyjne, takie jak stosunek mocy do wybiegu podczas ruchu i czas przebywania na stacji, w celu przywrócenia pociągu do zdefiniowanego dla niego przedziału czasowego. Nie ma maszynisty ani personelu przydzielonego do towarzyszenia pociągowi, co odpowiada GoA4.

W stopniu automatyzacji 4 systemy ATC działają w ramach ogólnego systemu sygnalizacji z blokadą, automatycznym nadzorem pociągu, wykrywaniem nie zajętości torów i funkcjami komunikacyjnymi.

Przykłady

Są dwa sposoby na wdrożenie zautomatyzowanych systemów kolejowych. Pierwszy wymusza budowę i eksploatację oddzielnych szlaków kolejowych. Tak funkcjonują np. autonomiczne kolejki na lotniskach we Frankfurcie czy też Singapurze. Ta działa już od 2003 roku. Dołączają do nich linie metra w poszczególnych miastach. Przykładowo w Paryżu jeszcze przed Igrzyskami Olimpijskimi w 2024 r. pociągi będą kursować po 200 km nowych, zautomatyzowanych torach wokół miasta, obsługujących 68 nowych stacji. Podobnie rozwija się Sydney, które może się poszczycić silną siecią transportu multimodalnego, łączącą autobusy, pociągi i promy. Obejmuje ona w pełni zautomatyzowany system metra, który został otwarty w 2019 roku i jest jednym z najbardziej innowacyjnych systemów komunikacji na świecie

Właściwie mogłaby do tych miast dołączyć Warszawa ze swoją linią metra M2, przygotowaną do jazdy autonomicznej, lecz nie ma to wciąż zgody Urzędu Transportu Kolejowego z powodu braku odpowiednich przepisów krajowych.

Drugi sposób zakłada, że pociągi autonomiczne współdzielą tory ze zwykłymi maszynami. W październiku 2021 roku. podczas Światowego Kongresu ITS (ang. *Intelligent Transport Systems*) w Hamburgu jeździły cztery zrobotyzowane pociągi na 23. kilometrowym odcinku linii S-Bahn S2 między stacjami Berliner Tor a Bergedorf.

Zdaniem Deutsche Bahn i Siemens, które wspólnie opracowały i zbudowały te pociągi, mogą one przewieźć do 30% więcej pasażerów, skrócić czas przejazdów, znacznie poprawić punktualność i zaoszczędzić ponad 30% energii elektrycznej. To przekonało Hamburg, aby od września 2022 roku robo-pociągi zaczęły normalnie kursować na tej trasie. W konsekwencji władze miasta chcą zrobotyzować całą sieć S-Bahn do 2030 roku. Ma ona 147 kilometrów długości i sześć linii, po których jeździ 200 pociągów, obsługujących 69 przystanków. Większość z nich uda się doposażyć w nową technologię cyfrową umożliwiającą jazdę autonomiczną, lecz oprócz tego zdecydowano się zakupić nowe 64 składy z najnowszej serii 490, które będą dostarczane sukcesywnie od 2025 roku.

Podstawą techniczną cyfrowych operacji kolejowych jest przyszła norma europejska ATO w połączeniu z europejskim systemem sterowania pociągiem ETCS (European Train Control System). Pociągi odbierają sygnały sterujące drogą radiową. Na wszelki wypadek maszynista ma dalej być na pokładzie.

POJAZDY PŁYWAJĄCE

Po trwającej 40 dni podróży i pokonaniu około 3500 mil morskich autonomiczny statek „Mayflower" (MAS400) dopłynął do portu Halifax w Nowej Szkocji w Ameryce Północnej 5 czerwca 2022 r. Po raz pierwszy w dziejach pokonał ocean statek bez człowieka na pokładzie, sterowany sztuczną inteligencją.

MAS400 został zaprojektowany i zbudowany przez organizację non-profit ProMare zajmującą się badaniami morskimi. Koncern IBM pełnił rolę wiodącego partnera technologicznego i naukowego, zaś jego technologie w zakresie automatyzacji, sztucznej inteligencji i obliczeń brzegowych zasilały system dowodzenia (AI) w kierowaniu statkiem i podejmowaniu decyzji w czasie rzeczywistym na morzu. Na pokładzie statku – obecnie cumującego z powrotem w Plymouth w Wielkiej Brytanii - znajduje się 6 kamer, ponad 30 czujników i 15 urządzeń edge, z których płyną dane do „kapitana AI" – centralnego systemu informatycznego, aby mógł je właściwie interpretować i analizować. Dzięki temu „kapitan AI" może przestrzegać prawa morskiego, podejmując kluczowe decyzje w ułamku sekundy, takie jak zmiana trasy wokół zagrożeń lub zwierząt morskich, a wszystko to bez interakcji albo interwencji człowieka.

Stopnie autonomiczności

Komitet ds. Bezpieczeństwa na Morzu Międzynarodowej Organizacji Morskiej (ang. IMO) przyjął w 2018 roku cztery stopnie automatyzacji statków oznaczanych odtąd skrótem MASS (ang. *Maritime Autonomous Surface Ships*).

- *Stopień pierwszy:* Statek ze zautomatyzowanymi procesami i wspomaganiem decyzji: Marynarze są na pokładzie, aby obsługiwać i kontrolować systemy i funkcje statku. Niektóre operacje mogą być zautomatyzowane i czasami nienadzorowane, ale z marynarzami na pokładzie gotowymi do przejęcia kontroli.
- *Stopień drugi:* Zdalnie sterowany statek z marynarzami na pokładzie: Statek jest kontrolowany i obsługiwany z innego miejsca. Marynarze są dostępni na pokładzie, aby przejąć kontrolę i obsługiwać systemy i funkcje pokładowe.
- *Stopień trzeci:* Zdalnie sterowany statek bez marynarzy na pokładzie: Statek jest kontrolowany i obsługiwany z innego miejsca. Na pokładzie nie ma marynarzy.

- *Stopień czwarty:* W pełni autonomiczny statek: System operacyjny statku jest w stanie samodzielnie podejmować decyzje i określać działania.

Odtąd działania IMO mają na celu ustanowienie spójnego systemu regulacyjnego dla MASS. Wynika to z konieczności, aby nadążać za już powstającymi projektami inteligentnych statków. To nie tylko MAS400, lecz np. pierwszy na świecie elektryczny i autonomiczny kontenerowiec Yara Birkeland, którego budowę rozpoczęto w 2017 r. Długość całkowita wynosi 80 metrów, szerokość całkowita zaś 15 metrów, natomiast maksymalne zanurzenie to 6 metrów. Statek ma pojemność 120 TEU i nośność 3200 ton. Napęd statku stanowią dwa pędniki azymutalne ze śrubami o średnicy 2,2 metra, napędzane silnikami elektrycznymi o mocy 900 kW każdy. Energię zapewniają zestawy akumulatorów litowo-jonowych o pojemności 6,8 MWh. Prędkość maksymalna wynosi 15 węzłów, podczas gdy ekonomiczna to 6–7 węzłów. Statek wyposażono w urządzenia do automatycznego cumowania.

Obecnie pływa w testowych rejsach pod norweską banderą i służy do przewożenia kontenerów ze sztucznymi nawozami z Porsgrunn do Breviku. Na razie jest sterowany konwencjonalnie, lecz po próbach przejdzie na tryb autonomiczny. Wtedy zostanie zdemontowany mostek, a kontrolę nad statkiem będzie sprawować centrum w Horten.

Armator chwali się, że dzięki przyjętym rozwiązaniom emisja CO2 spada rocznie o 1000 ton i zastępuje 40 tys. przejazdów samochodami ciężarowymi z silnikiem Diesla. Zapłacono za niego około 25 milionów dolarów, trzy razy więcej niż za konwencjonalny statek tej wielkości, lecz armator jest pewien, że zaoszczędzi do 90% kosztów eksploatacji. Zresztą rząd norweski wsparł budowę grantem w wysokości 16,7 miliona dolarów. W ślad za Norwegami podążyły firmy takie jak ABB, Wartsila i Rolls-Royce, które również zademonstrowały swoje projekty pilotażowe ze zdalnym pilotowaniem statków.

Ciekawe są kolejne ustalenia IMO na rzecz MASS. Uznano, że dla każdego instrumentu związanego z bezpieczeństwem i ochroną na morzu oraz dla każdego stopnia autonomii zostaną określone przepisy, które:

- mają stosować się do MASS i zapobiegać operacji MASS; lub
- mają zastosowanie do MASS i nie uniemożliwiają operacji MASS i nie wymagają żadnych działań; lub
- mają zastosowanie do MASS i nie uniemożliwiają operacji MASS, ale mogą wymagać zmiany lub wyjaśnienia i/lub mogą zawierać luki; lub nie mają zastosowania do operacji MASS.

Komitet bierze pod uwagę także przepisy kolizyjne, załadunek i stateczność, szkolenie marynarzy i rybaków, poszukiwanie i ratownictwo, pomiar tonażu, bezpieczne kontenery i przyrządy statków pasażerskich w handlu specjalnym. W kolejnych etapach zostanie określone jakie obecne przyrządy sprawdzą się na statkach autonomicznych, a jakie wymagają wymiany, biorąc pod uwagę m.in. element ludzki, technologię i czynniki operacyjne.

Równolegle działa Komitet Prawny IMO, który deliberuje nad m.in.:

- rolą i odpowiedzialnością kapitana;
- rolą i odpowiedzialnością zdalnego operatora;
- spójnymi definicjami/terminologią MASS, w tym przewozem towarów wymagających specjalnych certyfikatów.

Niektóre konkretne terminy prawne, takie jak pojęcia „winy", „niedbalstwa" i „zamiaru", wymagają szczegółowych rozważań w kontekście potencjalnych szkód możliwych do wyrządzenia przez technologię autonomiczną. Stąd narodził się pomysł nieobowiązkowego Kodeksu MASS, który wejdzie w życie w 2025 r. Instrument ma stanowić podstawę dla obowiązkowego Kodeksu MASS, którego spodziewane wejście w życie ma nastąpić 1 stycznia 2028 r. Kodeksu miałby zastosowanie do statków towarowych i raczej rozszerzyłby, a nie zastąpił istniejące przepisy IMO dotyczące bezpieczeństwa.

Kodeks będzie opierał się na analizie ryzyka. Jedno się nie zmieni: wszystkie odpowiednie komitety IMO zgadzają się co do zasady, że zostanie wyznaczony człowiek jako „kapitan", który będzie odpowiedzialny za eksploatację każdego statku MASS, niezależnie od poziomu jego autonomiczności. Osoba ta będzie pracować w wyznaczonym zdalnym centrum operacyjnym (ROC) i będzie podlegać nowym wymogom szkoleniowym oraz certyfikacji. Definicja i kwalifikacje roli „zdalnego operatora" będą częścią dalszej debaty.

Wyzwania społeczne

Prace nad morską technologią autonomiczną mają jeszcze jedno na celu – przeciwdziałać coraz większej barierze rozwojowej transportu morskiego, jaką jest brak chętnych do pracy. W raporcie „Seafarer Workforce Report" [12] szacuje się, że do 2026 roku będzie potrzebnych dodatkowo 89 510 oficerów do obsługi światowej floty handlowej. Obecnie pracuje w niej łącznie 1,89 mln marynarzy, którzy obsługują ponad 74 tys. statków na całym świecie. Obecnie brakuje 26 240 oficerów z odpowiednimi certyfikatami. Popyt na marynarzy w 2021 roku przewyższył podaż.

Ponadto, niektóre kategorie oficerów są szczególnie deficytowe. Brakuje oficerów z doświadczeniem technicznym na poziomie zarządzania, a w sektorze zbiornikowców i *offshore* odnotowano niedobór starszych oficerów w dziale pokładowym na poziomie zarządzania. Niewielkim pocieszeniem jest fakt, że w ciągu ostatnich pięciu lat branża żeglugowa poczyniła znaczne postępy w zakresie obniżenia wskaźnika rotacji oficerów z 8 proc. do 6 proc., zatrzymania wykwalifikowanych marynarzy oraz zwiększenia liczby lat ich służby na morzu. W porównaniu do szacunków z raportu z 2015 r., wzrósł średni wiek oficerów służących na poziomie zarządzania i poziomie operacyjnym. To jednak czynnik działający demobilizująco na armatorów zwykle nieskorych do inwestycji w nowe technologie.

Z drugiej strony rosną wymagania ekologiczne wobec statków. Łatwiej zaprogramować ich trasę i zdalnie je nadzorować na zatłoczonych szlakach morskich. „Kapitan AI" i jego „załoga" nie musi wypoczywać ani jeść, co przekłada się na zwiększenie przestrzeni ładunkowej. Statek może płynąć wolniej pod zdalnie sterowanymi żaglami, radykalnie zmniejszając zużycie paliwa. A takie argumenty trafiają do wyobraźni armatorów.

ROBOTYKA POWIETRZNA

Jak łatwo się domyślić, lotnictwo, w tym drony, też się doczekały własnej skali autonomii. Obejmuje ona pięć stopni od braku autonomii po pełną, tyle że stopień czwarty - wysokiej autonomii, w której AI samodzielnie pilotuje maszynę, dzieli się na trzy podklasy – im bardziej czujniki radzą sobie z wykrywaniem przeszkód i ludzi, tym jest lepiej. Należy jednak pamiętać o trzech sposobach rozumienia autonomii w przestrzeni powietrznej.

Współczesne samoloty pasażerskie właściwie mogłyby latać bez pilotów. Przepisy oraz troska o samopoczucie pasażerów wymagają jednak ich obecności na pokładzie. Startem zajmują się piloci, potem maszynę przejmuje autopilot, czasem aż do samego lądowania. Nie należy się spodziewać, że klasyczne lotnictwo całkowicie zrezygnuje z pilotów, mimo że loty z turystami na pokładzie autonomicznych rakiet kosmicznych i suborbitalnych maszyn jak Elona Muska, Jeffa Bezosa czy Richarda Bransona okazały się bezpieczne.

Z tego doświadczenia korzystają natomiast konstruktorzy autonomicznych taksówek jak np. VoloCity. To elektryczny dron pasażerski pionowego startu i lądowania, powstały w Badenii-Wirtembergii. Lata na wysokości 2 km z prędkością 70 km/h. Zaprojektowano go z myślą o podróżach lokalnych po obszarach miejskich do 30 kilometrów. Prawdopodobnie flota takich taksówek rozpocznie regularne loty w ciągu najbliższych dwóch-trzech lat.

Nad kabiną VoloCity rozpościera się ponad dziewięciometrowa obręcz z osiemnastoma wirnikami zasilanymi silnikami elektrycznymi. Wydzielany przez nie hałas na poziomie 65 decybeli jest mniejszy od zdecydowanej większości domowych odkurzaczy. Kabina mierzy trzy metry, jest szeroka na 1,2 metra i tyleż wysoka. Maszyna stoi na płozach. W środku kabiny zamontowano dwa, wygodne siedzenia jak w samochodzie i duży tablet, który służy do nie tylko rozrywki na pokładzie, lecz do śledzenia trasy przelotu na mapie i komunikacji z centrum zarządzania lotami. Przeszklony cały przód zapewnia świetną widoczność.

W powietrze wzbijają się też roboty, czyli drony naszpikowane różnego rodzaju czujnikami, poczynając od zwykłej kamery po laserowe skanery. Ich głównym zadaniem jest zbieranie danych i monitoring np. upraw (pozostawmy na boku zastosowania militarne). Nie szczędzi się sił i środków, by wprowadzić do miast autonomiczne drony transportujące różnego przesyłki np. medyczne. Zresztą, od lutego 2022 roku to już nie są testy, lecz codzienność szpitali w Sochaczewie i Pułtusku, które korzystają z usług dronów firmy LabAir (Farada Group), wysyłając nimi próbki do laboratoriów Alab w Warszawie.

Wyzwania prawne i technologiczne

By dalej mogła się rozwijać robotyka powietrzna, rządy muszą nadążyć z wdrożeniem odpowiedniego systemu prawnego i technologicznego. Te kwestie przeanalizowano w raporcie KPMG „Air Taxi Readiness Index 2022", czyli w indeksie gotowości do wprowadzenia zaawansowanej mobilności powietrznej [1].

Dokument ten ocenia przygotowanie 60 krajów, w tym Polski, do wprowadzenia nowej generacji pasażerskich pojazdów krótkiego lub pionowego startu i lądowania (S/VTOL). Polska znajduje się ok. 30 miejsca.

W dokumencie znajduje się pięć kategorii, według których każde państwo jest oceniane. Są to:

- Akceptacja konsumentów – ocena poziomu akceptacji społecznej.
- Infrastruktura – ocena gotowości infrastruktury do obsługi zaawansowanej mobilności powietrznej.
- Polityka i przepisy – ocena poziomu przygotowania polityki i przepisów do obsługi zaawansowanej mobilności powietrznej.
- Technologia i innowacje – ocena poziomu rozwoju technologii i innowacji związanych z zaawansowaną mobilnością powietrzną.
- Biznes – ocena poziomu gotowości biznesowej do obsługi zaawansowanej mobilności powietrznej.

Każda z tych kategorii składa się z kilku indywidualnych metryk, które są oceniane, a następnie łączone, aby stworzyć ogólny wynik dla kategorii i dla każdego państwa.

Skupmy się teraz na infrastrukturze. Rozległa infrastruktura jest niezbędna do wsparcia powszechnego użytkowania pojazdów S/VTOL, zwłaszcza w środowiskach miejskich. Chodzi o stacje ładowania energią elektryczną EV12, zasięg telefonii komórkowej 4G/5G, ilość ruchu lotniczego i szybkość połączenia mobilnego. Dodatkowo szalenie ważne są szerokopasmowy dostęp do Internetu, odpowiednie warunki klimatyczne (temperatura 18-30 stopni Celsjusza, słoneczne dni z niskim poziomem opadów i wilgotności) oraz gęstość wieżowców (jako wskaźnik miejsca lądowania w centrum miasta).

W tej kategorii liderem są Stany Zjednoczone, które przesunęły się na pierwsze miejsce w 2022 roku. Chiny spadły zaś na drugie miejsce, a po nich uplasowały się Korea Południowa, Zjednoczone Emiraty Arabskie i Japonia. Polska zajmuje 30 miejsce [1].

Wydawałoby się, że wysokie miejsca winny zajmować kraje Unii Europejskiej, tymczasem one plasują się w środku stawki. Dlaczego? EASA prezentuje zachowawczą i niezrozumiałą politykę wobec robotyki powietrznej. Można odnieść wrażenie, że drony przeszkadzają lotnictwu i najchętniej EASA zakazałaby ich lotów. Być może wpływa na to lobbing wielkich koncernów lotniczych i motoryzacyjnych, które nastawiają się na sprzedaż autonomicznych taksówek powietrznych. Dla nich każdy latający robot w przestrzeni do 120 metrów jest zagrożeniem.

A jak jeszcze byłoby tego mało, to EASA wprowadziła zaporowe ceny na homologację dronów transportowych, porównywalną do kosztów homologacji dużych samolotów pasażerskich. To luddyzm i obrona status quo.

PODSUMOWANIE

Na podstawie zebranych informacji można się pokusić o stwierdzenie, że przełom w zakresie implementacji transportu autonomicznego wydarzy się w ciągu najbliższych sześciu lat, z czego kluczowe będą lata 2027–2030. Połączą się wtedy w jeden gigatrend dotychczasowe megatrendy jak

energetyka oparta na wodorze, rozwój transmisji danych w technologii 5G, zwiększona pojemność baterii, dominacja OZE w miksie energetycznym oraz właśnie transport autonomiczny. Nie mamy czasu na rozważania, czy warto, tylko jak się do tego przygotować, by na tym skorzystać.

Zacznijmy od akceptacji i zaufania społecznego. Konieczne jest przekonanie ludzi o korzyściach płynących z transportu autonomicznego, ale także rozwianie obaw związanych z utratą miejsc pracy kierowców, prywatnością danych oraz możliwością nadużyć lub awarii systemów. Ponadto należy opracować ramy prawne dotyczące odpowiedzialności za wypadki, zasady dotyczące zgody na korzystanie z danych osobowych oraz przepisy dotyczące testowania i certyfikacji pojazdów autonomicznych.

Wiąże się z tym konieczność modyfikacji istniejącej infrastruktury drogowej, morskiej i powietrznej. W przypadku transportu lądowego należy zainstalować czujniki drogowe, inteligentne sygnalizacje świetlne i systemy komunikacji, aby umożliwić skuteczną interakcję pojazdów autonomicznych z otoczeniem.

Ponadto technologia autonomiczna obejmuje szereg zaawansowanych systemów, takich jak czujniki, algorytmy sztucznej inteligencji, systemy wizyjne, systemy nawigacji i wiele innych. Integracja i synchronizacja tych różnych systemów stanowią wyzwanie technologiczne, ponieważ wymaga ich skutecznego działania wraz z minimalnymi błędami. Przed wprowadzeniem transportu autonomicznego konieczne jest zaś przeprowadzenie rozległych testów i certyfikacji systemów autonomicznych.

Ważne jest także, aby zapewnić równy dostęp do transportu autonomicznego dla wszystkich grup społecznych. Konieczne jest uniknięcie sytuacji, w których tylko pewne obszary lub grupy społeczne mają korzystać z zalet tej technologii, podczas gdy inne są wykluczone.

LITERATURA

[1] Collie Brian, Decker Jordan, Fishman Jamie, Sridhara Raj, Wegscheider Augustin K., Wiesinger Michael. 2022. „Mapping the Future of Autonomous Trucking". *Boston Consulting Group & Kodiak Robotics.*

[2] J.D. Power. 2022. „U.S. Mobility Confidence Index (MCI) Study. In collaboration with Partners for Automated Vehicle Education and MIT Advanced Vehicle Technology Consortium". *J.D. Power.*

[3] Kosieliński Sławomir. 2022. „Projekt badawczy `„ROBOTnik: inkubator kwalifikacji robotycznych.". Transferhub,. *Fundacja Inicjatyw Społeczno-Ekonomicznych.*

[4] Kosieliński Sławomir. 2022. „Roboty drogowe". *Polityka* nr 1/2 (3345): 95– – 97.

[5] Kosieliński Sławomir. 2023. „Na prawo patrz". Burza. *Rynek dronów w Polsce.* Raport Fundacji Instytut Mikromakro i Łukasiewicz – Instytutu Lotnictwa. Edycja 2023, : 32– - 37. *Fundacja Instytut Mikromakro.*

[6] KPMG. 2018. „Autonomous Vehicles Readiness Index. Assessing countries' openness and preparedness for autonomous vehicles". 2018. *KPMG.*

[7] KPMG. 2022. „Air Taxi Readiness Index 2022. Assessing the preparedness of 60 territories in the race for Advanced Air Mobility". *KPMG.*

[8] Napiórkowski Marcin. 2022. „Naprawić przyszłość. Dlaczego potrzebujemy lepszych opowieści, żeby uratować świat". *Wydawnictwo Literackie.*

[9] National Highway Traffic Safety Administration. 2021. „Automated Vehicles for Safety". *www.nhtsa.gov/technology-innovation/automated-vehicles-safety.*

[10] Renub Research. 2022. „Global Autonomous Vehicles Market, Size, Forecast 2022-2030, Industry Trends, Growth, Impact of COVID-19, Opportunity Company Analysis". 2022. *Renub Research.*

[11] Rutkowska Joanna, Walewski Mateusz, Wroński Maciej. 2016. „Rynek pracy kierowców w Polsce". *PwC.*

[12] Seafarer Workforce Report. 2021. „Baltic and International Maritime Council – BIMCO & International Chamber of Shipping". *Seafarer Workforce Report.*

Publikuj artykuły

www.przegladtelekomunikacyjny.pl

WSPÓŁCZESNE ARCHITEKTURY SYSTEMÓW OPARTYCH O MIKROSERWISY

CONTEMPORARY ARCHITECTURES OF MICROSERVICES-BASED SYSTEMS

Andrzej Matiolański[1];

[1] AGH Akademia Górniczo-Hutnicza, Kraków, andrzej.matiolanski@agh.edu.pl

DOI: 10.15199/59.2023.4.6

Streszczenie: **Architektura mikroserwisów to podejście do tworzenia aplikacji jako zbioru luźno powiązanych, niezależnych, drobnych usług, które wspólnie tworzą kompletną aplikację. W artykule przedstawiono trzy ewolucyjne podejścia do architektury systemu opartego na mikroserwisach. Dla typowej architektury dokonano eksperymentalnej ewaluacji parametrów użytkowych na przykładzie czasu odpowiedzi na zapytania. W artykule przedstawiono także syntetyczne wnioski oraz kierunku rozwoju tego rodzaju rozwiązań.**

Abstract: **Microservices architecture is an approach to building applications as a collection of loosely coupled, independent, small services that collectively form a complete application. The article presents three evolutionary approaches to microservices-based system architecture. For a typical architecture, an experimental evaluation of performance parameters was conducted, focusing on response time to queries. The article also presents synthetic conclusions and directions for the development of such solutions.**

Słowa kluczowe: **aplikacje internetowe, architektura, czas odpowiedzi, mikroserwisy, wydajność**

Keywords: **architecture, microservices, performance, response time, web applications**

1. Wstęp

Architektura mikroserwisów dla SaaS (Software as a Service) to podejście do tworzenia aplikacji, które polega na dekompozycji oprogramowania na niezależne, skalowalne i elastyczne moduły, nazywane mikroserwisami. Każdy mikroserwis odpowiada za jedną funkcjonalność i może być rozwijany, wdrażany oraz skalowany niezależnie od innych. Architektura ta jest szczególnie przydatna w przypadku aplikacji SaaS, gdyż pozwala na łatwe zarządzanie, rozbudowę i skalowanie systemu.

Architektura oparta o mikroserwisy zyskała na popularności w ostatnich latach, głównie ze względu na swoją elastyczność, skalowalność i odporność na awarie. Przykłady zastosowań tej architektury można znaleźć w różnych branżach i w wielu znanych firmach takich jak NetFlix, AirBnB, Amazon, Spotify [1, 2]. Warto zauważyć, że nie tylko duże korporacje korzystają z architektury mikroserwisów. Ta architektura jest stosowana również przez mniejsze firmy i startupy, które chcą zapewnić swoim klientom szybkie, niezawodne i elastyczne usługi. Architektura mikroserwisów sprawdza się szczególnie w przypadku złożonych systemów, które wymagają łatwej rozbudowy, aktualizacji oraz integracji z innymi usługami [3, 4].

2. Budowa i charakterystyka

Architektura systemów teleinformatycznych oparta o mikroserwisy składa się z trzech kluczowych komponentów [5]:

- Mikroserwisy: Niezależne, małe aplikacje odpowiedzialne za jedną funkcjonalność, posiadające własne bazy danych, API (interfejs) i mechanizmy uwierzytelniania. Mikroserwisy są rozwijane, wdrażane i skalowane niezależnie, co pozwala na większą elastyczność systemu.

- Brama API (API Gateway): Jest to punkt wejścia do systemu, który zarządza żądaniami klientów i kieruje je do odpowiednich mikroserwisów. Brama API może również obsługiwać uwierzytelnianie, autoryzację, limitowanie ruchu oraz agregację danych z różnych mikroserwisów.

- Komunikacja między mikroserwisami: Aby zapewnić sprawną współpracę pomiędzy mikroserwisami, stosuje się różne protokoły komunikacyjne, takie jak REST, gRPC, GraphQL czy asynchroniczne techniki komunikacji, np. wykorzystujące brokery wiadomości (Kafka, RabbitMQ).

Cechy jakimi charakteryzuje się architektur oparta o mikroserwisy to [6, 7]:

- Skalowanie: Architektura mikroserwisów pozwala na horyzontalne skalowanie poszczególnych elementów systemu, co umożliwia szybkie dostosowanie się do zmiennych potrzeb klientów i zasobów infrastruktury.

- Odporność na awarie: Mikroserwisy są projektowane tak, aby działać niezależnie od innych, co minimalizuje ryzyko wpływu awarii jednego z nich na pozostałe elementy systemu.

- Ciągła integracja / Ciągłe dostarczanie (CI/CD): W architekturze mikroserwisów dla SaaS, często stosuje się automatyzację procesów wdrożeniowych oraz ciągłą integrację i dostarczanie oprogramowania, co pozwala na szybsze wprowadzanie zmian oraz łatwiejsze zarządzanie systemem.

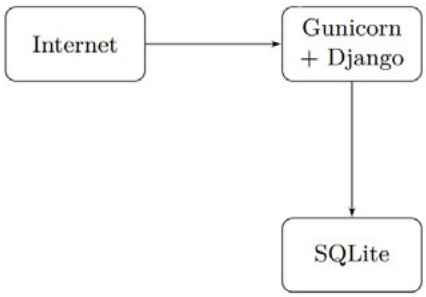
- Monitoring i logowanie: Aby skutecznie zarządzać systemem opartym na mikroserwisach, niezbędne jest gromadzenie, analiza i monitorowanie danych dotyczących wydajności, błędów i innych parametrów poszczególnych mikroserwisów. Popularne narzędzia do tego celu to m.in. Prometheus, Grafana, ELK Stack czy Jaeger.

2.1. Rozważane przykłady architektu

Dla potrzeb przyszłych eksperymentów wykorzystano serwis internetowy udostępniający pliki zbudowany w języku Python z frameworkiem Django. Będzie on uruchamiany za pomocą Gunicorn czyli serwera WSGI HTTP dla Pythona. WSGI, czyli Web Server Gateway Interface, to standard interfejsu między serwerami internetowymi a aplikacjami Pythona. Gunicorn działa jako pośrednik między serwerem internetowym, taki jak Nginx czy Apache, a aplikacją Pythona. Serwer internetowy przyjmuje żądania HTTP od użytkowników, przekazuje je do Gunicorna, który następnie obsługuje żądanie za pomocą aplikacji WSGI. Gunicorn zwraca odpowiedź do serwera internetowego, który z kolei zwraca ją do użytkownika.

2.2. Podstawowa architektura mikroserwisów

Najprostsza architektura mikroserwisów może zostać uruchomiona lokalnie i de facto od swojego monolitycznego odpowiednika różni się jedynie strukturą. Składa się z modułu Bramy API i mikroserwisu (np. Gunicorn + Django) oraz bazy danych (np. SQLite). Brama API jest połączona z mikroserwisem. Z uwagi na ułomność takiej architektury jest ona zazwyczaj wykorzystywana jedynie do celów testowych. W naszym przypadku posłuży jako referencja dla dalszej analizy.

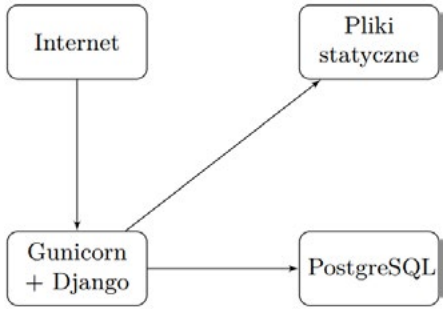


Rysunek 1: Podstawowa architektura mikroserwisów

2.3. Typowa architektura mikroserwisów

W najprostszej formie architektury mikroserwisów, może ona składać się tylko z kilku instancji mikroserwisów, które komunikują się ze sobą przez RESTful API, a do koordynacji ruchu używają API Gateway. Chociaż ta architektura jest prostsza niż bardziej złożone warianty, nadal oferuje wiele korzyści, takich jak łatwość zarządzania, niezależność wdrażania i skalowania jest jednak podatna na awarie.

W naszym przypadku architektura składa się z Gunicorn i Django, a dodatkowe mikroserwisy oferują funkcjonalność bazy danych i przestrzeni dyskowej (w formie statycznych plików). W tym przypadku jako baza danych zdecydowanie lepiej sprawdzi się PostgreSQL, do którego połączenie może odbywać się za pomocą protokołu HTTP.



Rysunek 2: Typowa architektura mikroserwisów

W tym przypadku mamy już do czynienia z kompletną architekturą. Baza danych oraz przestrzeń dyskowa jako serwisy mogą zostać powielone w razie potrzeby, aby uniknąć tzw. wąskich gardeł przetwarzania. W tym przypadku wydajność spada w przypadku przesycenia zapytaniami Gunicorn+Django.

2.4. Równoważenie obciążenia i multiplikacja

Modyfikacją typowej architektury jest powielenie niektórych mikroserwisów. W opisanym przypadku pojawiły się redundantne instancje Gunicorn + Django, które mogą przetwarzać zapytania. Nie komunikują się one pomiędzy sobą, ale posiadają taki sam dostęp do pozostałych zasobów systemu. Takie podejście niesie ze sobą szereg korzyści w aspektach takich jak:

- Niezawodność: Redundancja polega na posiadaniu wielu kopii tego samego mikroserwisu, co sprawia, że cały system jest bardziej odporny na awarie pojedynczych węzłów. Jeśli jeden z węzłów ulegnie awarii, pozostałe kopie mogą przejąć jego obowiązki, zapewniając ciągłość działania.
- Odporność na błędy: W architekturze z redundancją, błędy wprowadzone do jednej kopii mikroserwisu nie wpłyną na cały system. Pozwala to na szybsze wykrywanie i naprawianie błędów, a także na lepszą izolację problemów.
- Bezpieczeństwo: Redundancja może poprawić bezpieczeństwo systemu, ponieważ utrudnia atakowanie wszystkich instancji mikroserwisu jednocześnie. W przypadku ataku, pozostałe kopie mogą nadal funkcjonować i utrzymywać działanie systemu.

Nie byłoby to jednak możliwe bez zastosowania modułu równoważenia obciążenia (ang. Load Balancer), którego zadaniem jest równomierne obciążenia zasobów, a co za tym idzie wzrost elastyczności. W naszym przypadku wykorzystano równoważenie obciążenia w oparciu o Nginx, który działa jako odwrócony serwer proxy (reverse proxy). Dystrybuuje ruch sieciowy do różnych mikroserwisów w celu równoważenia obciążenia i zapewnienia wysokiej dostępności. Jego działanie składa się z następujących etapów:

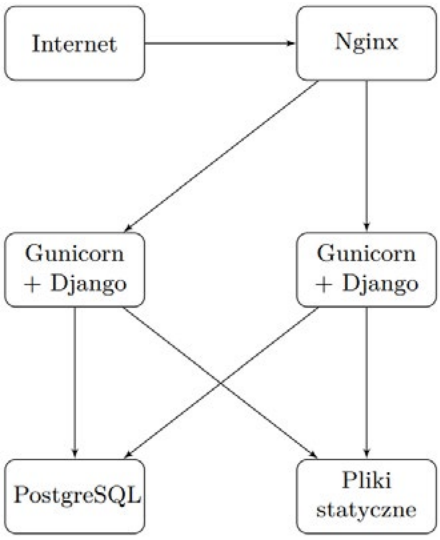
1. Przychodzące zapytania: Kiedy klient (np. przeglądarka internetowa) wysyła zapytanie do aplikacji,

zapytanie to trafia najpierw w celu równoważenia obciążenia do Nginx.

2. Algorytm równoważenia obciążenia: Nginx obsługuje różne algorytmy równoważenia obciążenia, takie jak round-robin, least connections, czy IP hash. Algorytm jest odpowiedzialny za wybór jednego z serwerów backendowych do obsługi każdego przychodzącego zapytania. Wybór serwera zależy od konfiguracji i obciążenia poszczególnych serwerów.

3. Przekierowanie zapytania: Po wyborze odpowiedniego serwera backendowego, Nginx równoważy obciążenie i przekierowuje zapytanie do tego serwera. Serwer przetwarza zapytanie i generuje odpowiedź.

4. Odpowiedź do klienta: Odpowiedź od serwera backendowego jest przekazywana z powrotem przez Nginx do klienta, który wysłał zapytanie.

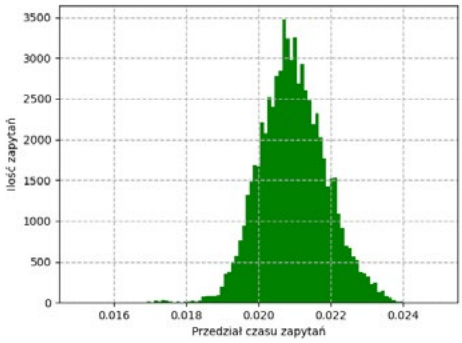


Rysunek 3: Architektura mikroserwisów z równoważeniem obciążenia

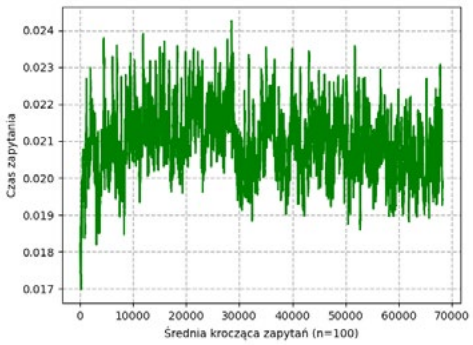
3. Eksperyment i wyniki

W ramach dalszej ewaluacji architektur mikroserwisów wykonano eksperyment polegający na pomiarze czasów odpowiedzi na zapytania. W tym celu wykorzystano typową architekturę mikroserwisów opisaną w poprzedniej sekcji artykułu. Zmierzonymi parametrami był czas potrzebny na komunikację pomiędzy mikroserwisami, czas wykonania zadania polegającego na prostym obliczeniu numerycznym oraz suma obu tych wartości. W tym celu wykonano 60 000 zapytań. Wyniki zostały przedstawione w postaci histogramów odzwierciedlających rozkład zmierzonych czasów oraz na wykresach przedstawiających średnią kroczącą zmierzonych czasów dla okna czasowego wynoszącego 100 próbek.

Zaobserwowanych wyników (rys. 4) można wywnioskować że dla krótkich zadań czas komunikacji nie może zostać pominięty. Jego optymalizacja może okazać się kluczowa w przypadku gdy obsługujemy istotnie dużą liczbę krótkich zapytań. Także w przypadku czasu komunikacji zdarzają się okresowe szczyty (rys. 5) które powodują wydłużenie łącznego czasu przetwarzania zapytania. Jest to szczególnie istotne w systemach działających w czasie rzeczywistym, kiedy to nadmierne opóźnienia są niepożądane.

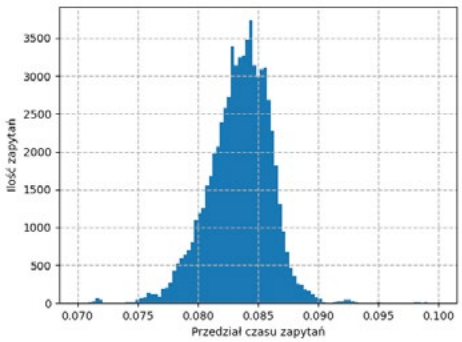


Rysunek 4: Histogram czasu komunikacji w sekundach

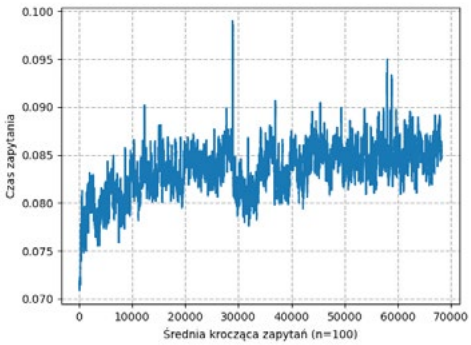


Rysunek 5: Zmierzony czas potrzebny na komunikację w sekundach

W przypadku czasu przetwarzania zauważamy początkowy wzrost aż do ustalenia stałej wartości która nieznacznie fluktuuje (rys. 7). Tu także mamy do czynienia z fluktuacją, jednak nie przekracza ona 15%. Analizując łączny czas zauważamy obecność niewielkiej ilości czasów poza typowym rozkładem. Powodem jest prawdopodobnie odpowiednie nałożenie na siebie fluktuacji wynikającej z czasów komunikacji oraz przetwarzania (rys. 6). W tabeli 1. przedstawiono wartości średnie minimalne maksymalne oraz medianę zmierzonych parametrów.



Rysunek 6: Histogram czasu komunikacji w sekundach



Rysunek 7: Zmierzony czas przetwarzania w sekundach

Tabela 1: Porównanie czasów odpowiedzi w sekundach w sekundach

Parametr	Komunikacja	Przetwarzanie	Łącznie
Średnia	0,0209	0,0834	0,1041
Minimum	0,0129	0,0535	0,0713
Maksimum	0,2387	0,7634	0,7812
Mediana	0,0203	0,0803	0,1004

4. Podsumowanie

Architektura oparta na mikroserwisach zdobyła dużą popularność w ostatnich latach ze względu na korzyści, takie jak modularność, niezależne wdrażanie i łatwość skalowania. W przyszłości można spodziewać się dalszego rozwoju tej architektury oraz ewolucji technologii związanych z mikroserwisami. Oto kilka trendów, które mogą wpłynąć na przyszłość architektur opartych o mikroserwisy [8, 9]:

- Serverless: Serverless to podejście do budowania i wdrażania aplikacji, w którym zarządzanie infrastrukturą jest całkowicie ukryte przed deweloperem. W architekturach serverless, mikroserwisy są uruchamiane jako funkcje, które są wywoływane na żądanie, dzięki czemu deweloperzy mogą skupić się na pisaniu kodu, a nie na zarządzaniu infrastrukturą.
- Service Mesh: Service Mesh to dedykowana warstwa infrastruktury, która ułatwia zarządzanie komunikacją między mikroserwisami. Takie rozwiązania, jak Istio, Linkerd czy Consul, pomagają w automatyzacji i bezpieczeństwie komunikacji między mikroserwisami, monitoringu, zarządzaniu ruchem sieciowym, czy obsłudze awarii.
- Event-driven architecture: W architekturach zorientowanych na zdarzenia, mikroserwisy komunikują się za pomocą zdarzeń i reagują na nie, zamiast bezpośrednio wywoływać inne serwisy. Event-driven architecture może prowadzić do wzrostu niezawodności i elastyczności.
- Orkiestracja i zarządzanie kontenerami: Technologie kontenerów, takie jak Docker, stały się kluczowe dla wdrażania i zarządzania mikroserwisami.
- Większe zastosowanie sztucznej inteligencji i uczenia maszynowego: W przyszłości architektury oparte na mikroserwisach mogą coraz częściej wykorzystywać sztuczną inteligencję i uczenie maszynowe, aby automatycznie optymalizować działanie systemu, zarządzać zasobami i wykrywać problemy.

Wraz z rozwojem technologii i narzędzi wspierających, architektury oparte na mikroserwisach prawdopodobnie będą nadal ewoluować, dostarczając jeszcze bardziej elastycznych, skalowalnych i efektywnych rozwiązań dla szerokiego spektrum zastosowań.

Literatura

[1] Newman, S. (2015). *Building Microservices: Designing Fine-Grained Systems*. O'Reilly Media. ISBN: 978-1491950357.

[2] Richardson, C. (2018). *Microservices Patterns: With examples in Java*. Manning Publications. ISBN: 978--1617294549.

[3] Newman, S. (2014). From Monolith to Microservices: Evolutionary Patterns to Transform Your Monolith. InfoQ. Retrieved from `https://www.infoq.com/articles/monolith-to-microservices-sam-newman/`

[4] Nadareishvili, I., Mitra, R., McLarty, M., Amundsen, M. (2016). *Microservice Architecture: Aligning Principles, Practices, and Culture*. O'Reilly Media. ISBN: 978-1491956250.

[5] Fowler, M. (2014). Microservices: a definition of this new architectural term. Retrieved from `https://martinfowler.com/articles/microservices.html`

[6] Pahl, C., & Jamshidi, P. (2016). Microservices: A systematic mapping study. In *Proceedings of the 6th International Conference on Cloud Computing and Services Science* (pp. 137-146).

[7] Thönes, J. (2015). Microservices. *IEEE Software*, 32(1), 116-116.

[8] F. Wang and J. Zhang, "Research on the Current Situation and Future Trend of Microservice Technology Development," *2022 IEEE 6th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Beijing, China, 2022, pp. 44--54, doi: 10.1109/IAEAC54830.2022.9929766.

[9] R. M. Munaf, J. Ahmed, F. Khakwani and T. Rana, "Microservices Architecture: Challenges and Proposed Conceptual Design," *2019 International Conference on Communication Technologies (ComTech)*, Rawalpindi, Pakistan, 2019, pp. 82-87, doi: 10.1109/COMTECH.2019.8737831.

SYSTEM DO MONITOROWANIA NARAŻENIA NA WIBRACJE
VIBRATION EXPOSURE MONITORING SYSTEM

Kazimierz Roman[1]; Jerzy Kołakowski[2]
[1] Instytut Radioelektroniki i Technik Multimedialnych, Politechnika Warszawska, Warszawa, kazimierz.roman.stud@pw.edu.pl
[2] Instytut Radioelektroniki i Technik Multimedialnych, Politechnika Warszawska, Warszawa, jerzy.kolakowski@pw.edu.pl

DOI: 10.15199/59.2023.4.7

Streszczenie: **Wibracje mogą być przyczyną poważnych schorzeń układu nerwowego krwionośnego i kostno-szkieletowego. Z tego powodu ich detekcja, określenie poziomu i czasu trwania są istotne na stanowiskach pracy narażonych na wibracje. W referacie opisano budowę systemu umożliwiającego rejestrację parametrów wibracji oraz transmisję alertów w przypadku przekroczenia zadanych poziomów i czasów. W artykule zamieszczono opis konstrukcji urządzeń monitorujących oraz wstępne wyniki badań.**
Abstract: **Vibrations can cause serious diseases of the circulatory and skeletal nervous systems. Therefore, their detection, determination of level, and duration are important in workplaces exposed to vibrations. The paper presents the design of a system for the recording of vibration parameters and the transmission of alerts if the predefined levels and times are exceeded. The article describes the design of devices and preliminary test results.**

Słowa kluczowe: **detekcja wibracji, IoT, BLE.**

Keywords: **vibration detection, IoT, BLE.**

1. WSTĘP

Wibracje, czyli drgania, są obecne w wielu dziedzinach naszego życia. Często są one skutkiem pracy maszyn wykorzystywanych na budowie. Niestety, wibracje mogą być przyczyną poważnych schorzeń układu nerwowego, krwionośnego oraz kostno-szkieletowego. W związku z tym ich detekcja, pomiar oraz monitorowanie są istotne dla osób narażonych na wibracje. Minimalne wymagania w zakresie ochrony zdrowia i bezpieczeństwa dotyczące narażenia pracowników na ryzyko spowodowane wibracjami zostały określone w Dyrektywie 2002/44/WE Parlamentu Europejskiego i Rady [1].

W referacie przedstawiono system monitorowania narażenia na wibracje w formie noszonej etykiety, który umożliwia rejestrację parametrów wibracyjnych oraz transmisję alertów za pomocą modułu radiowego w przypadku przekroczenia określonych poziomów parametrów wibracji. Jednym z założeń projektu jest fakt, że urządzenie ma być wykorzystywane w trudnym środowisku - np. na budowie. Wymusza to ograniczenia konstrukcyjne rejestratora takie jak miniaturyzacja, mała masa, odporność na pyły oraz wodę. Środowisko stawia wysokie wymagania również dla modułu radiowego, który musi zapewnić stabilną łączność z centrum systemu monitorowania w trudnych warunkach propagacyjnych.

Podczas opracowywania systemu skorzystano ze standardu ISO-5349 [2] dotyczącego oceny narażenia pracowników na wibracje ręcznych narzędzi roboczych oraz innych urządzeń. Określa on metodykę pomiarową oraz kryteria oceny narażenia na wibracje. Metodyka badań wpływu wibracji na całe ciało została opisana w normie ISO 2631-1 [3].

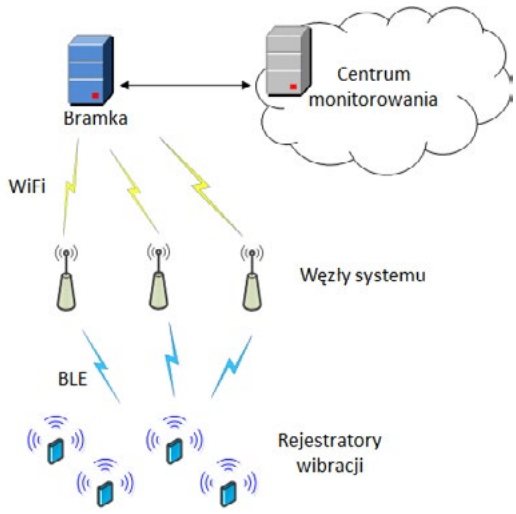
W literaturze można znaleźć różne przykłady realizacji podobnych systemów do pomiaru wibracji. W [4] autorzy przedstawili system do pomiaru wibracji w formie opaski na dłoń, która może być umieszczona pod rękawicą ochronną stosowaną przez operatorów urządzeń budowlanych. Urządzenie to jest połączone z "jednostką bazową", przymocowaną do pasa operatora, która może przechowywać dane, generować alarmy i przesyłać je za pomocą interfejsu radiowego.

W [5] autorzy zaproponowali podobny system do tego z artykułu [4], jednak „jednostka bazowa", została umieszczona w większej odległości od czujnika inercyjnego. System uwzględniał możliwość przetwarzania i monitorowania danych pochodzących z wielu modułów pomiarowych noszonych przez różnych pracowników. Nieco inne podejście do badań wibracji, polegające na pomiarze za pomocą czujników umieszczanych bezpośrednio na źródle drgań zaprezentowano w [6] i [7]. Opisano w nich metodologię oraz wyniki pomiarów wibracji kierownicy traktora i motocykla.

2. KONCEPCJA SYSTEMU

2.1. Ogólna architektura systemu

Architekturę systemu do monitorowania narażenia na wibracje przedstawiono na rys. 1. Użytkownicy systemu wykorzystują noszone (naręczne) rejestratory umożliwiające detekcję wibracji i transmisję komunikatów alarmowych w przypadku przekroczenia założonych poziomów i czasów narażeń. Dodatkowo urządzenia wysyłają cyklicznie pakiety umożliwiające ich lokalizację. Transmisje z rejestratorów są odbierane przez sieć węzłów, które pełnią jednocześnie rolę węzłów systemu lokalizacyjnego. Na podstawie zarejestrowanych poziomów sygnałów w bramce jest określana lokalizacja rejestratora. Wykorzystanie większej liczby węzłów umożliwia zwiększenie prawdopodobieństwa odbioru komunikatów w trudnym środowisku propagacyjnym. Bramka wysyła informację o zdarzeniach oraz dane lokalizacyjne do centrum monitorowania.

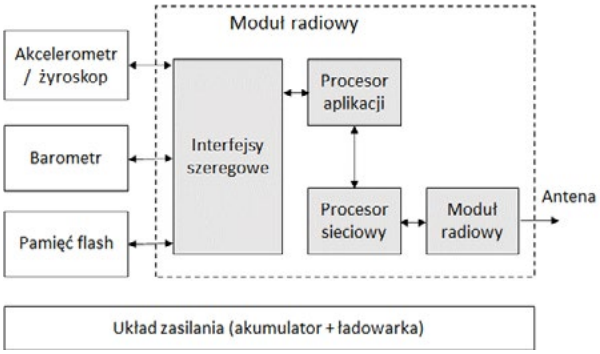


Rys. 1. Architektura systemu

Rejestratory wibracji wykorzystują do transmisji komunikatów interfejs BLE (Bluetooth Low Energy), komunikacja węzłów z bramką odbywa się za pomocą sieci WiFi. Oprócz detekcji wibracji system wyposażono w funkcje umożliwiające detekcję upadku oraz detekcję bezruchu przez zadany czas.

2.2. Układ do rejestracji wibracji

Na rys. 2 przedstawiono schemat blokowy układu rejestratora wibracji.



Rys. 2. Schemat blokowy układu rejestratora

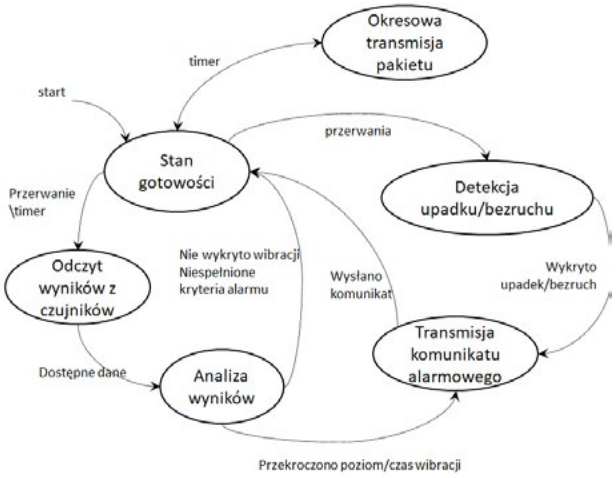
Pracą układu steruje mikrokontroler nRF5340, w którym zastosowano dwa procesory ARM Cortex M33. Jeden z nich pełni rolę procesora aplikacji, z kolei drugi obsługuje stos protokołów łącza radiowego. Układ umożliwia realizację łączności w paśmie 2,4 GHz zgodnie ze standardami BLE, ZigBee i Thread. W układzie wykorzystano gotowy moduł BL5340 PA firmy Laird zawierający oprócz mikrokontrolera zestaw wzmacniaczy. Pozwalają one na pracę z mocą wyjściową do 18 dBm, która nie jest osiągalna w zwykłych modułach BLE oraz zwiększają czułość modułu radiowego do -108 dBm. Obecność oddzielnego, wydajnego procesora aplikacji daje możliwość implementacji złożonych algorytmów przetwarzania danych bez wpływu na działanie łącza radiowego.

W urządzeniu wykorzystano układ BMI270 firmy Bosch Sensortec zawierający akcelerometr i żyroskop. Dodatkowym czujnikiem, w który wyposażono układ jest barometr BMP390. Dane z czujników są wykorzystywane do detekcji upadku. Wykorzystana w układzie dodatkowa pamięć flash umożliwia zapis informacji o zdarzeniach i służy do celów diagnostycznych.

2.3. Oprogramowanie rejestratora

Wykorzystanie czujnika inercyjnego BMI270 firmy Bosh Sensortec wymagało napisania oprogramowania umożliwiającego użycie zaawansowanych funkcji oferowanych przez czujnik. Czujnik posiada dwa wyprowadzenia INT1/INT2, które podłączone do mikrokontrolera pozwalają na generację przerwań. W nocie katalogowej [3] producent określił kilka kluczowych dla projektu przerwań: „anymotion" – generowane w chwili wykrycia ruchu (wibracji), „nomotion" – generowane w chwili ustania ruchu oraz „FIFO full" – generowane, gdy kolejka FIFO przechowująca wyniki pomiarów się zapełni. Dzięki nim mikrokontroler może pozostać w stanie uśpienia gdy wibracje nie przekraczają pewnego oraz wybudzić się w odpowiednim momencie, aby rozpocząć akwizycję oraz przetwarzanie danych z kolejki FIFO. Układ BMI270 skonfigurowano tak, że na linii INT1 przekazywane są przerwania dotyczące ruchu (motion, nomotion), a na linii INT2 przerwania dotyczące kolejki FIFO (FIFO full). Komunikacja z czujnikiem odbywa się z wykorzystaniem interfejsu SPI.

Diagram stanów programu sterującego pracą rejestratora przedstawiono na rys. 3. Po włączeniu zasilania układ przebywa w stanie gotowości zapewniającym minimalizację zużytej energii. Okresowo timer wymusza transmisję pakietów do celów lokalizacyjnych.



Rys. 3. Diagram stanów programu rejestratora

Odczyt wyników z czujników następuje po przerwaniu zgłoszonym przez kolejkę FIFO lub w wyniku zadziałania timera określającego częstość odbioru wyników z barometru. Następnie układ przechodzi do stanu Analiza wyników, w którym są przeprowadzane obliczenia umożliwiające detekcję wibracji, określenie poziomu i czasu trwania drgań. W przypadku spełnienia kryterium alarmu układ przechodzi do stanu transmisji komunikatu alarmowego. Detekcja upadku i bezruchu jest zrealizowana za pomocą przerwań generowanych przez układ BMI270. Pozwalają one na zgłoszenie przerwania po zadanym czasie bezruchu, przekroczeniu zadanej wartości przyspie-

szenia, wykryciu swobodnego spadku. Dodatkowym kryterium detekcji upadku jest względna zmiana wysokości określana na podstawie wskazań barometru.

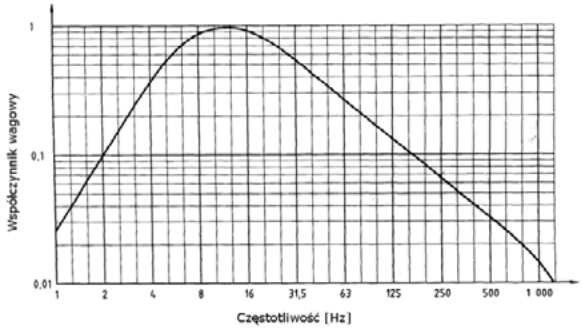
Oprogramowanie rejestratora zostało napisane w języku C z wykorzystaniem systemu czasu rzeczywistego Zephyr. Funkcje DSP umożliwiające wykonywanie zoptymalizowanych obliczeń na procesorach ARM zostały dostarczone przez bibliotekę CMSIS-DSP.

2.4. Przetwarzanie zarejestrowanych sygnałów

Według standardu ISO 5349-1 wartość skuteczna (r.m.s) przyspieszenia a_{hv} jest określona wzorem:

$$a_{hv} = \sqrt{a_{hwx}^2 + a_{hwy}^2 + a_{hwz}^2} \quad (1)$$

gdzie $a_{hwx}, a_{hwy}, a_{hwz}$ są wartościami skutecznymi przyspieszenia zmierzonymi dla każdej z trzech osi, przefiltrowanymi za pomocą filtru o charakterystyce przedstawionej na rys. 4. Do wyznaczenia wartości a_{hw} dla każdej z osi wykorzystano analizę widmową (FFT).



Rys. 4. Charakterystyka częstotliwościowa filtru [2]

Gdy obliczone są wartości skuteczne przyspieszenia dla kilku wibracji a_{hv} i zarejestrowane czasy T trwania każdej z wibracji można przystąpić do wyznaczenia dziennej ekspozycji na wibrację $A(8)$, która odnosi się do 8-godzinnego dnia pracy. Jest ona określona wzorem:

$$A(8) = \sqrt{\frac{1}{T_0}\sum_{i=1}^{n} a_{hvi}^2 \; T_i} \quad (2)$$

gdzie a_{hvi} to wartość skuteczna przyspieszenia w przedziale czasu T_i, n to liczba okresów ekspozycji na wibrację, T_0 – czas pracy (8h)

Dla przykładu, jeśli założymy, że w ciągu dnia pracy zarejestrowaliśmy trzy wibracje, dla których wartości a_{hv} wyniosły kolejno $2\,m/s^2$, $3.5\,m/s^2$, $10\,m/s^2$, i które trwały kolejno 1h, 3h, 0.5h to dzienna ekspozycja na wibracje wyniesie:

$$A(8) = \sqrt{\frac{1}{8h}\left[\left(2\frac{m}{s^2}\right)^2 * 1h + \left(3{,}5\frac{m}{s^2}\right)^2 * 3h + \left(10\frac{m}{s^2}\right)^2 * 0{,}5h\right]}$$

$$= 3{,}4\frac{m}{s^2} \quad (3)$$

Dla wibracji przekazywanych na kończyny górne w dyrektywie [1] podano „dzienną wartość działania narażenia" (2,5 m/s^2) określającą próg, po którego przekroczeniu ryzyko uszczerbku na zdrowiu spowodowane wibracjami znacząco rośnie. Drugą wartością ustaloną w dyrektywie [1] jest „maksymalny dopuszczalny poziom narażenia na wibracje (5 m/s^2). Obie wartości odnoszą się do równania (2).

Istotną kwestią konieczną do uzyskania poprawnych wyników wydaje się być usunięcie z pomiarów akcelerometru czynnika związanego ze stałym przyspieszeniem ziemskim. Aby tego dokonać należałoby odjąć wektor przyspieszenia ziemskiego od każdego pomiaru. Problem w tym, że ziemski układ odniesienia różni się od układu odniesienia akcelerometru umieszczonego w rejestratorze. Jednym z rozwiązań jest wykorzystanie macierzy obrotu bądź struktur algebraicznych – kwaternionów, w celu obrotu wektora zmierzonego przyspieszenia w danej próbce tak, aby znajdował się on w ziemskim układzie odniesienia, a następnie odjęcie wektora przyspieszenia ziemskiego. Takie podejście wymaga precyzyjnej znajomości kąta obrotu dla każdej z osi. Nie jest to trywialne zadanie, jednak okazuje się, że korzystając z danych dostarczanych przez żyroskop oraz akcelerometr można skonstruować filtr Kalmana, który umożliwi precyzyjne wyznaczenie kąta obrotu dla poszczególnych osi. Dalsze prace nad projektem uwzględniają analizę oraz testy tego zagadnienia.

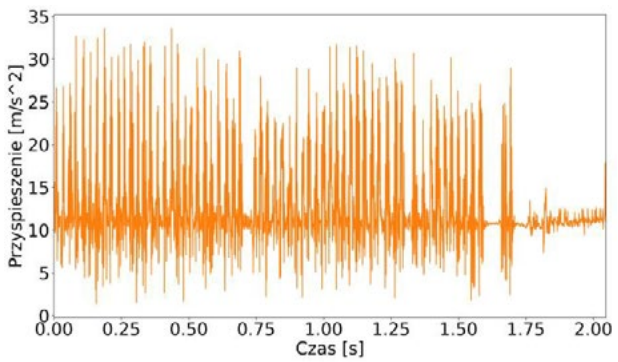
3. BADANIA UKŁADU

3.1. Układ eksperymentalny

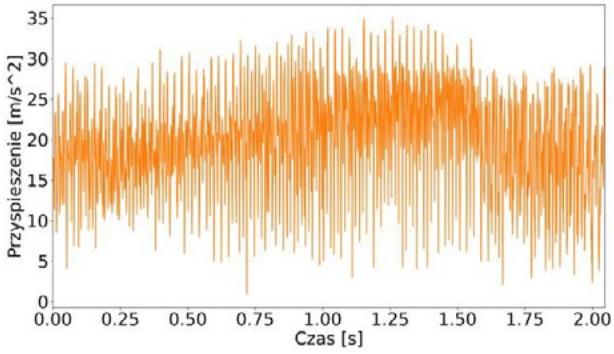
Układ eksperymentalny stanowiący podstawę do wstępnych badań możliwości pomiarowych urządzenia prezentował się następująco: jeden rejestrator wibracji został umieszczony na nadgarstku, drugi zaś został przymocowany bezpośrednio do urządzenia. Pomiar wibracji został dokonany podczas użytkowania typowych urządzeń: szlifierki oscylacyjnej oraz wiertarki z funkcją udaru. Częstotliwość próbkowania wynosiła $1000\,Hz$, a pojedynczy pomiar trwał $2s$.

3.2. Wstępne wyniki badań

Na rys. 5 i 6 zaprezentowano przykładowe wykresy zmian przyspieszenia zarejestrowane za pomocą układu umieszczonego na nadgarstku podczas korzystania ze szlifierki i wiertarki z udarem. Wartości zarejestrowanych przyspieszeń są zbliżone.

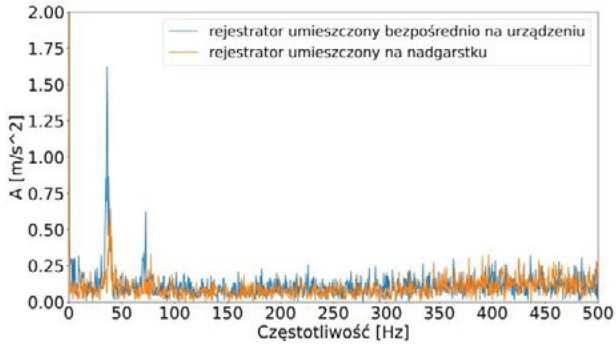


Rys. 5. Wykres zmian przyspieszenia – szlifierka oscylacyjna - rejestrator umieszczony na nadgarstku

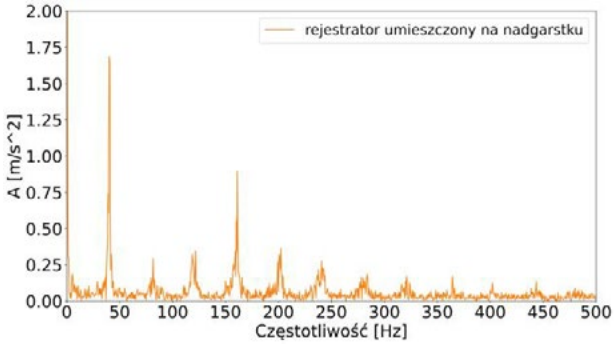


Rys. 6. Wykres zmian przyspieszenia – wiertarka (udar) - rejestrator umieszczony na nadgarstku

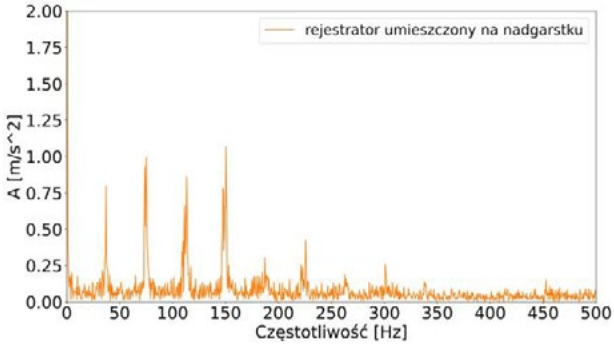
Wyniki analizy widmowej zarejestrowanych zmian przyspieszenia przedstawiono na rys. 7, 8 i 9. Na podstawie wykresów można stwierdzić, że widma zarejestrowanych drgań są znacząco różne. W przypadku wszystkich urządzeń zaobserwowano różnice związane z lokalizacją czujnika przyspieszenia.



Rys. 7. Widmo wibracji – szlifierka oscylacyjna



Rys. 8. Widmo wibracji – wiertarka z wyłączonym udarem



Rys. 9. Widmo wibracji – wiertarka z włączonym udarem

4. PODSUMOWANIE

W referacie przedstawiono koncepcję systemu do monitorowania wibracji w środowisku budowlanym oraz wstępne wyniki badań prototypowego układu rejestratora. Wykorzystanie w układzie zaawansowanych mikrokontrolerów udostępniających dużą moc obliczeniową i funkcje transmisji radiowej pozwala na miniaturyzację urządzenia i realizację obliczeń w czujniku, co z kolei umożliwia redukcję ilości danych transmitowanych w interfejsie radiowym.

Wstępne wyniki badań zrealizowanego układu potwierdzają skuteczność jego działania. Badania wykazały, że na amplitudę rejestrowanych drgań ma wpływ miejsce i sposób umocowania urządzenia (np. amplituda drgań rejestrowanych na nadgarstku jest znacząco mniejsza niż w przypadku pomiaru bezpośrednio na obudowie urządzenia).

W referacie wykorzystano oprogramowanie zrealizowane w ramach projektu finansowanego przez Narodowe Centrum Badań i Rozwoju (program EraPerMed nr umowy PerMed/II/34/PerHeart/2022).

LITERATURA

[1] Dyrektywa 2002/44/WE Parlamentu Europejskiego i Rady z dnia 25 czerwca 2002 r. w sprawie minimalnych wymagań w zakresie ochrony zdrowia i bezpieczeństwa dotyczących narażenia pracowników na ryzyko spowodowane czynnikami fizycznymi (wibracji) (szesnasta dyrektywa szczegółowa w rozumieniu art. 16 ust. 1 dyrektywy 89/391/EWG)

[2] ISO 5349-1:2001 Mechanical vibration — Measurement and evaluation of human exposure to hand-transmitted vibration — Part 1: General requirements

[3] ISO 2631-1:1997 Mechanical vibration and shock — Evaluation of human exposure to whole-body vibration — Part 1: General requirements

[4] T. Addabbo et al., "A wearable Low-cost Measurement System for Estimation of Human Exposure to Vibrations," 2019 IEEE 5th International forum on Research and Technology for Society and Industry (RTSI), Florence, Italy, 2019, pp. 442-446,

[5] R. Morello, C. De Capua and A. Meduri, "A Wireless Measurement System for Estimation of Human Exposure to Vibration During the Use of Handheld Percussion Machines," in IEEE Transactions on Instrumentation and Measurement, vol. 59, no. 10, pp. 2513-2521, Oct. 2010,

[6] Phulsagar Amruta Savta “A Study of Reduction in the Vibrations of Steering Wheel of Agricultural Tractor” ISSN : 2248-9622, Vol. 6, Issue 10, (Part - 4) October 2016, pp.80-87

[7] Shivakumara BS, Sridhar V “Study of vibration and its effect on health of the motorcycle rider” Mangalore, South India : ISSN 0972-5997, Volume 9, Issue 2; Apr-Jun 2010

[8] BMI270 datasheet, Bosch Sensortec, rev.1.5, March 2023

EMULACJA ŚRODOWISKA DLA ZASTOSOWANIA PROTOKOŁU IN-BAND NETWORK TELEMETRY

ENVIRONMENT EMULATION FOR IN-BAND NETWORK TELEMETRY

Sylwester Kaczmarek[1]; Mateusz Krasoń[2]

[1] Politechnika Gdańska, Wydział ETI, Gdańsk, kasyl@eti.pg.edu.pl
[2] Politechnika Gdańska, Wydział ETI, Gdańsk, m.krason01@wp.pl

DOI: 10.15199/59.2023.4.8

Streszczenie: **Określenie jakości obsługi strumieni pakietów w sieci przełączników wymaga odpowiedniego środowiska badawczego w którym prowadzi się doświadczenia i pomiary wybranych wielkości. Protokół In-band Network Telemetry jest jednym z narzędzi, które można wykorzystać do realizacji tych zadań. W pracy zaproponowano zwirtualizowane środowisko badawcze w którym można emulować sieć przełączników programowalnych w języku P4 wraz z implementacją tego protokołu. Przeprowadzone testy wykazały poprawność działania zaproponowanego rozwiązania.**

Abstract: **Determining the quality of packet stream service in a switch network requires an appropriate research environment in which experiments and measurements of selected quantities are carried out. In-band Network Telemetry is one of the tools that can be used to accomplish these tasks. The paper proposes a virtualized research environment in which a network of programmable switches in the P4 language can be emulated along with the implementation of this protocol. The tests carried out showed the correct operation of the proposed solution.**

Słowa kluczowe: **emulacja, język P4, Mininet, protokół INT, przełącznik bmv2.**

Keywords: **bmv2 switch, emulation, INT protocol, Mininet, P4 language.**

1. WSTĘP

Ciągły wzrost zapotrzebowania na usługi sieciowe spełniające określone wymagania jakościowe, na bazie których są lub mają być realizowane różnorodne usługi dostarczane klientom, stwarza konieczność wprowadzenia mechanizmów umożliwiających bieżącą i szybką ocenę parametrów określających jakość. Jednocześnie oczekuje się, że będą one na tyle uniwersalne i rozwojowe aby mogły obejmować możliwie duży zakres przypadków zastosowań. Należy nadmienić, że dla przypadku strumieni RTP do tego celu mógł być wykorzystany protokół RTCP, który pracuje w trybie out-of-band [4]. Ciągły postęp w rozwoju technologii sieciowych oraz możliwość programowania elementów sieciowych na poziomie układów obsługujących strumienie pakietów doprowadził do opracowania protokołu In-band Network Telemetry (INT), który spełnia wyartykułowane ale również szersze oczekiwania [12]. Istota tego rozwiązania sprowadza się w ogólności do wstrzyknięcia w odpowiednie miejsce pakietu klina o określonej strukturze pól, który razem z pakietem przemierza na drodze połączeniowej elementy, które mogą być zaprogramowane w języku P4 [9,10]. W efekcie ich działania na tych polach mogą być wykonane określone operacje wynikające z aktualnego stanu elementu sieci i potrzeb funkcji sterowania oraz nadzoru. Pakiet taki po przejściu między wybranymi punktami (źródłem i odbiorcą) zawiera w klinie z protokołu INT dane, które mogą być przetworzone i dostarczać informacji o bieżących parametrach jakościowych na przebytej drodze. Zauważmy że skojarzenie protokołu INT i języka P4 stwarza tu znacznie szersze możliwości zastosowania niż by to wynikało z samego protokołu INT. Mamy tu do czynienia z efektem synergii obu rozwiązań.

Przebadanie różnych możliwości tego typu rozwiązań wymaga wytworzenia odpowiedniego środowiska badawczego. Jednym z przykładów jest rozwiązanie o nazwie „Deep Insight" zaprezentowane przez firmę Intel [3]. Podstawowym zadaniem tego rozwiązania jest testowanie i nadzorowanie przepływu pakietów w czasie rzeczywistym. Między innymi jest to wykrywanie błędów tj. utraconych pakietów, bądź zbyt dużych opóźnień podczas przepływu pakietów przez urządzenia programowalne oraz realizowane w sieci funkcje. Innym rozwiązaniem jest „Minicom" [7] w którym jest możliwość generowania strumieni pakietów z klinem protokołu INT oraz skierowania ich do rzeczywistych lub emulowanych programowalnych przełączników, które mogą na polach tego klina wykonywać wcześniej zaprogramowane przetwarzanie.

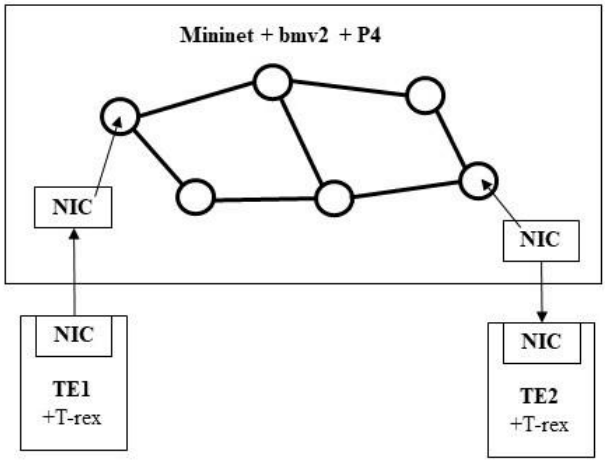
W niniejszej pracy opisano zaproponowane i zrealizowane własne środowisko testowe dla protokołu INT wraz z możliwością zastosowania języka programowania P4, jako narzędzia do prowadzenia szerszego zakresu badań jakości usług, ich sterowania i nadzoru. Jednocześnie chcąc się uniezależnić od rozwiązań sprzętowych postanowiono, że będzie ono w całości zwirtualizowane.

W związku z tym chcąc przekazać informacje na ten temat w dalszej części pracy w rozdziale drugim opisano architekturę i wymagania funkcjonalne tego środowiska, w rozdziale trzecim opisano jego realizację, natomiast w rozdziale czwartym scharakteryzowano wybrane przeprowadzone testy. Pracę kończy podsumowanie.

2. ARCHITEKTURA I WYMAGANIA FUNKCJONALNE NA ŚRODOWISKO

Na poziomie sieci węzłów architektura sprowadza się do sieci przełączników połączonych łączami a do prze-

łączników dołączone są terminale TE. Terminale te generują strumienie pakietów, które są obsługiwane przez przełączniki zgodnie z zainstalowanymi w nich mechanizmami. Jednocześnie przełączniki te są programowalne w języku P4 i dzięki temu można programować te mechanizmy. Umożliwia to im przetwarzanie pól nagłówków pakietów wygenerowanych w terminalach. Z kolei terminale mają możliwość wstawiania klina do pakietu zgodnie ze standardem protokołu INT. Ogólna architektura środowiska badawczego została przedstawiona na rys. 1.



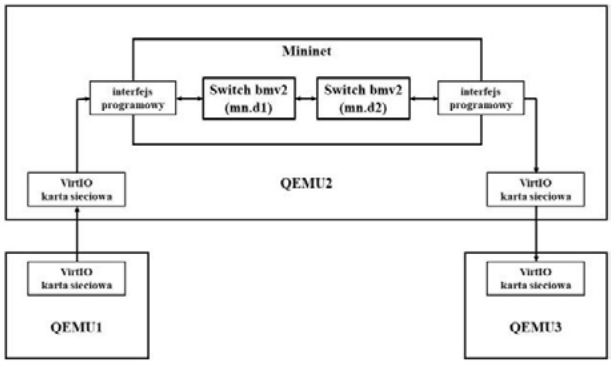
Rys. 1. Ogólna architektura środowiska badawczego

Dla jego realizacji konieczne jest oprogramowanie umożliwiające utworzenie sieci przełączników i tego celu wykorzystano środowisko Mininet [8] wraz z przełącznikami bmv2 (behavioral model version 2) [1], które mogą być programowalne przy pomocy pliku *json* wygenerowanego przez kompilator języka P4. Z kolei połączenia między elementami architektury wykonano przy pomocy kart NIC (Network Interfejs Controller) wykorzystując oprogramowanie VirtIO. Generacja i odbiór pakietów odpowiednio w terminalach TE1 i TE2 został zrealizowany w oparciu o oprogramowanie T-rex [2].

Skonsolidowanie tego oprogramowania, przy założeniu wirtualizacji całego środowiska, wymagało zrealizowania dodatkowych zadań, które zostaną opisane w kolejnym rozdziale traktującym o realizacji tego środowiska. Te dodatkowe zadania to nie tylko czynności związanie z odpowiednim konfigurowaniem wymienionego oprogramowania ale także wytworzenie własnego oprogramowania, np. interfejsów programowych między elementami, czy też generowanie pakietów z ramkami protokołu INT.

3. REALIZACJA ŚRODOWISKA EMULACYJNEGO

Realizacja podstawowego założenia, że środowisko badawcze w całości ma być zwirtualizowane wymagało wykorzystania trzech maszyn wirtualnych zbudowanych w oparciu o oprogramowanie QEMU (Quick EMUlator) [13]. Są nimi maszyny o nazwach QEMU1, QEMU2 i QEMU3 z oprogramowaniem Linux Ubuntu 18.04 Bionic [14]. Ta wersja systemu gwarantowała kompatybilność z każdym wymaganym w niniejszym projekcie oprogramowaniem. W maszynie QEMU1 został umieszczony terminal TE1 generujący strumień pakietów (oprogramowanie T-rex) wraz z polami protokołu INT oraz wirtualna karta sieciowa NIC na bazie programu VirtIO. Terminal odbiorczy TE2 został umieszczony w maszynie QEMU3 z podobnym oprogramowaniem jak w terminalu TE1. Z kolei sieć przełączników została umieszczona w maszynie QEMU2 bazując na oprogramowaniu środowiska Mininet oraz wersji przełącznika bmv2 charakteryzującego się możliwością programowania w języku P4. Przełączniki bmv2 wykorzystywane przez oprogramowanie Mininet zostały skonfigurowane jako kontenery dockerowe i zostały wykorzystane ich dwie instancje mn.d1 i mn.d2. Umieszczono tu również karty sieciowe na bazie programu VirtIO umożliwiające podłączenie obu terminali. Dla celów testowania realizacji koncepcji tego projektu sieć przełączników składa się tylko z dwóch połączonych ze sobą przełączników bmv2 co nie stanowi istotnego ograniczenia dla testowania założonej funkcjonalności. W celu zrealizowania połączenia kart sieciowych ze przełącznikami bmv2 wirtualizowanymi przy pomocy oprogramowania Mininet konieczne było zrealizowanie interfejsu programowego oraz zautomatyzowanie generowania przestrzeni adresowej przez Mininet. Struktura tak zrealizowanego środowiska została pokazana na rys. 2.



Rys. 2. Struktura zrealizowanego wirtualnego środowiska badawczego

W celu utworzenia połączenia między wirtualną kartą NIC a przełącznikiem utworzonym w Mininet zrealizowano interfejs programowy wykorzystując do tego celu następujące oprogramowanie: VirtIO – wirtualizowana karta sieciowa, nadawanie i odbiór pakietów; Veth – wirtualny Ethernet dla łączenia dwóch urządzeń; *forward_scrtx.py* – programy napisane w języku Python nasłuchujące na odpowiednich interfejsach i przekazujące nasłuchane pakiety do zdefiniowanego interfejsu. W projekcie zainstalowano dwa interfejsy programowe, jeden między kartą NIC a przełącznikiem bmv2 (mn.d1), a drugi między przełącznikiem bmv2 (mn.d2) a kartą NIC. W celu uruchomienia przełączników mn.d1 i mn.d2 oraz ustanowienia odpowiedniej konfiguracji między nimi został napisany program w języku Python dla środowiska Mininet o nazwie *mininet_script.py*.

Wytworzone środowisko ma służyć badaniom zastosowania protokołu INT to istnieje konieczność generowania strumienia pakietów zawierających ramki protokołu INT. Ponieważ standard opisuje szereg trybów pracy to w projekcie wybrano tryb INToverUDP z nagłówkiem MD.

W tym rozwiązaniu wstrzykiwanie ramek INT ma miejsce między UDP header a UDP payload i składa się ono z INT Shim header (klin), INT-MD Metadata Header, które mają określoną strukturę i zawartość [11]. W związku z tym w projekcie konieczne było zaimplementowanie rozwiązania, które w terminalu TE1 będzie wstrzykiwało ramkę protokołu INT. Można wyróżnić tu cztery sposoby realizacji tego zadania, z których dwa zostały zrealizowane w projekcie. Tymi sposobami są mianowicie:

- napisanie oprogramowania w języku Python wstrzykującego odpowiednie struktury do pakietów UDP,

- wykorzystanie „gotowych" pakietów UDP ze wstrzykniętym INT-MD i metadanymi,

- wstrzykiwanie ramek przy pomocy P4 w sprzęcie, bądź w emulowanych przełącznikach bmv2,

- wykorzystanie oprogramowania host-int firmy Intel w celu wstrzykiwania ramek INT-MD do pakietów UDP.

Z uwagi na poczynione założenia co do sposobu realizacji środowiska badawczego w projekcie aktualnie można zastosować dwa pierwsze sposoby. W przypadku pierwszego sposobu wykorzystano w terminalu TE1 następujące oprogramowanie: T-rex – generator ruchu pakietów UDP firmy CISCO, Veth – wirtualne urządzenie Ethernetowe, *int_inject.py* – napisany program w języku Python, który wstrzykuje ramki INT do pakietów UDP. Drugi sposób korzysta z wcześniej wygenerowanych pakietów UDP z ramkami protokołu INT. Ramki UDP zostały wygenerowane przy pomocy oprogramowania T-rex, do których następnie przy pomocy języka P4 zaimplementowanego na fizycznych przełącznikach wstrzyknięto ramki INT-MD po nagłówku UDP. Z uwagi na ograniczenia sprzętowe w projekcie skorzystano z plików *.pcap* udostępnionych przez autora pracy [5]. Zarówno w jednym jak i w drugim sposobie generacji wymagana jest odpowiednia konfiguracja oprogramowania T-rex.

Kolejnym bardzo ważnym zadaniem jakie musi być realizowane w wytworzonym środowisku jest możliwość programowania przełączników w języku programowania P4. W tym celu wykorzystano przełącznik bmv2 z modelem o nazwie *v1model* znajdujący się w bibliotece *v1model.p4* [1]. Zgodnie z ogólnymi założeniami przełączników z językiem programowania P4 każdy przełącznik bmv2 składa się z bloków o następujących nazwach: *parser*, *checksum verification*, *ingress processing* (Match-Action), *egress processing* (Match-Action), *checksum computation* (Update) i *deparser*. W ramach zrealizowanego projektu napisano program w języku P4 [10], który realizuje każdy z wyżej wymienionych bloków. Są to odpowiednio: *parser*, *Ingress processing*, *checksum computation* oraz *deparser*. Pozostałe bloki zostały pominięte ze względu na to, że bloki te nie realizują żadnych zadań, które mogłyby mieć wpływ na działanie modelu. Niemniej zostały one uwzględnione w kodzie programu ale posiadają puste komendy *apply{}*. Oczywiście gdyby trzeba było dla poszczególnych bloków zmienić zadania w procesie przetwarzania pakietu to należy to odpowiednio oprogramować w języku programowania P4. Po skompilowaniu i wygenerowaniu pliku *json* tym plikiem zaprogramować przełącznik. Dla tu przedstawionego przypadku program w języku P4 nosi nazwę *p4_program.p4*.

Tak przygotowane elementy środowiska badawczego wymagały skonsolidowania, odpowiedniego skonfigurowania i przeprowadzenia testów. Z uwagi na narzuconą objętość artykułu zostało to w miarę możliwości opisane w kolejnym rozdziale.

4. TESTY ŚRODOWISKA

Konsolidacja w jedną całość wyspecyfikowanego ogólnie dostępnego wolnego i wytworzonego oprogramowania wymagała w pierwszej kolejności instalacji trzech maszyn wirtualnych QEMU1, QEMU2 i QEMU3 [13]. Polecenie uruchomienia oprogramowania maszyny QEMU1 pokazano na rys. 3. Podobnie jest to przeprowadzane dla maszyny QEMU3. W przypadku maszyny QEMU2 polecenie jest nieco bardziej złożone z uwagi na potrzebę wykorzystania, aż czterech instancji NIC – dwie w celu komunikacji QEMU1-QEMU2 i dwie w celu komunikacji QEMU2-QEMU3. Ważna jest kolejność uruchamiania maszyn a po ich uruchomieniu aktywacja wirtualnych kart sieciowych i odpowiednia do potrzeb konfiguracja.

```
qemu-system-x86_64 \
-enable-kvm -nographic -serial mon:stdio \
-name user1-vm1  -smp 3 -m 8G  \
-hda bionic-server-cloudimg-amd64.img  \
-net user,hostfwd=tcp::8005-:22   \
-nic user,model=virtio-net-pci  \
-netdev socket,id=sock1,listen=localhost:6667  \
-device virtio-net-pci,netdev=sock1  \
-machine q35
```

Rys. 3. Polecenie uruchomienia maszyny QEMU1

Na tym etapie realizacji środowiska bardzo ważne było przetestowanie działania poszczególnych interfejsów pomiędzy maszynami oraz wewnątrz maszyny QEMU2, tzn. interfejsu TE1 (NIC) z NIC w Mininet, interfejsu programowego, interfejsu TE2 (NIC) z NIC w Mininet, drugiego interfejsu programowego. Wykorzystano do tego celu *ping* oraz *tcpdump*. Na rys. 4 pokazano wynik działania *ping* na QEMU1, a na rys. 5 pokazano nasłuch przy pomocy *tcpdump* na QEMU2. Ostatecznie przetestowano połączenie między maszynami QEMU1 a QEMU3.

```
root@krasny-vm1:~# ping -I enp0s3 10.0.70.70
PING 10.0.70.70 (10.0.70.70) from 10.0.69.69 enp0s3: 56(84) bytes of data.
64 bytes from 10.0.70.70: icmp_seq=1 ttl=64 time=6.18 ms
64 bytes from 10.0.70.70: icmp_seq=2 ttl=64 time=3.13 ms
64 bytes from 10.0.70.70: icmp_seq=3 ttl=64 time=3.39 ms
64 bytes from 10.0.70.70: icmp_seq=4 ttl=64 time=2.75 ms
64 bytes from 10.0.70.70: icmp_seq=5 ttl=64 time=3.84 ms
64 bytes from 10.0.70.70: icmp_seq=6 ttl=64 time=3.10 ms
64 bytes from 10.0.70.70: icmp_seq=7 ttl=64 time=2.59 ms
64 bytes from 10.0.70.70: icmp_seq=8 ttl=64 time=14.7 ms
64 bytes from 10.0.70.70: icmp_seq=9 ttl=64 time=3.02 ms
64 bytes from 10.0.70.70: icmp_seq=10 ttl=64 time=11.4 ms
```

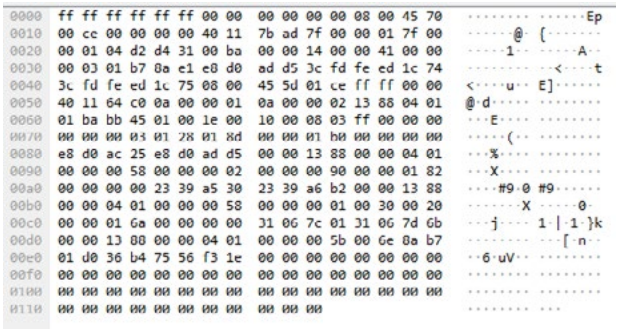
Rys. 4. Wynik działania ping na QEMU1

```
root@krasny-vm1:~# tcpdump -i enp0s5
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on enp0s5, link-type EN10MB (Ethernet), capture size 262144 bytes
20:49:24.793147 IP 10.0.69.69 > 10.0.70.70: ICMP echo request, id 2490, seq 7, length 64
20:49:24.794281 IP 10.0.70.70 > 10.0.69.69: ICMP echo reply, id 2490, seq 7, length 64
20:49:25.794721 IP 10.0.69.69 > 10.0.70.70: ICMP echo request, id 2490, seq 8, length 64
20:49:25.795656 IP 10.0.70.70 > 10.0.69.69: ICMP echo reply, id 2490, seq 8, length 64
20:49:26.796584 IP 10.0.69.69 > 10.0.70.70: ICMP echo request, id 2490, seq 9, length 64
20:49:26.797420 IP 10.0.70.70 > 10.0.69.69: ICMP echo reply, id 2490, seq 9, length 64
20:49:27.798238 IP 10.0.69.69 > 10.0.70.70: ICMP echo request, id 2490, seq 10, length 64
20:49:27.798987 IP 10.0.70.70 > 10.0.69.69: ICMP echo reply, id 2490, seq 10, length 64
20:49:28.798897 IP 10.0.69.69 > 10.0.70.70: ICMP echo request, id 2490, seq 11, length 64
20:49:28.799305 IP 10.0.70.70 > 10.0.69.69: ICMP echo reply, id 2490, seq 11, length 64
20:49:29.803853 IP 10.0.69.69 > 10.0.70.70: ICMP echo request, id 2490, seq 12, length 64
20:49:29.805232 IP 10.0.70.70 > 10.0.69.69: ICMP echo reply, id 2490, seq 12, length 64
```

Rys. 5. Nasłuch na QEMU2

Przed przystąpieniem do całościowego testu konieczne było skonfigurowanie generatora T-rex aby generował pakiety UDP oraz umożliwiał wstrzykiwanie ramek protokołu INT, według sposobów opisanych w poprzednim rozdziale 3.

Test wytworzonego środowiska polegał na wygenerowaniu w TE1 pakietów UDP wraz z ramką protokołu INT, przesłanie ich przez sieć do TE2 i obejrzenie jego zawartości oraz skonfrontowanie z wynikiem działania zaprogramowanym dla przełączników bmv2 w Mininet. Wyniki były zgodne z oczekiwaniami i na rys. 6 pokazano zawartość bajtową jednego takiego pakietu odebranego w TE2.



Rys. 6. Zawartość bajtowa odebranego pakietu w TE2

Przeprowadzone testy zrealizowanego środowiska wykazały, że osiągnięto zamierzony cel i środowisko to można wykorzystać do szeregu badań z wykorzystaniem ramek protokołu INT. Jednocześnie środowisko to może być dalej wzbogacane o nowe elementy funkcjonalne w zależności od potrzeb, tym bardziej, że jest ono w całości zwirtualizowane.

5. PODSUMOWANIE

W pracy wykazano, że istnieje możliwość zrealizowania zwirtualizowanego środowiska badawczego w którym daje się zintegrować protokół INT oraz język programowania P4. W tym zwirtualizowanym środowisku dzięki wykorzystaniu oprogramowania Mininet oraz przełącznika bmv2 można emulować sieć przełączników w której obsługiwany jest protokół INT. Do budowy tego środowiska wykorzystano otwarte oprogramowanie uzupełniając je własnym niezbędnym oprogramowaniem umożliwiającym konsolidację wszystkich elementów oprogramowania. Testy koncepcji przeprowadzono dla najprostszej konfiguracji funkcjonalno-sieciowej. Pozytywny wynik testów umożliwia zarówno dalszy rozwój tego środowiska badawczego jak i prowadzenie szczegółowych badań obsługi strumieni pakietów. W pierwszym przypadku dotyczy on zaimplementowania pozostałych trybów pracy protokołu INT oraz zwiększenie złożoności badanych sieci bazujących na Mininet oraz przełącznikach bmv2 z możliwością ich programowania w języku P4. Badania będą skupione w kierunku przyjmowania do obsługi i samej obsługi przepływów charakteryzujących się zróżnicowaniem wymaganej jakości obsługi. To z kolei wiąże się z koniecznością montowania w przełącznikach w języku P4 różnych mechanizmów ich obsługi. Mechanizmy te mogą wykorzystywać rzeczywiste dane, dotyczące jakości obsługi, pozyskiwane z pomiarów w sieci [5]. Do przeprowadzania pomiarów bieżących parametrów jakościowych można właśnie wykorzystać protokół INT, który między innymi został do tego zaprojektowany. Tym samym w zrealizowanym środowisku badawczym można będzie weryfikować algorytmy przyjmowania klas strumieni pakietów do obsługi oraz ich obsługi z różnicowaniem jakości, uwzględniając na bieżąco informacje o dynamicznie zmieniającym się stanie sieci i osiąganych parametrach jakościowych. Umożliwi to także badanie algorytmów pod kątem optymalizacji wykorzystania zasobów.

LITERATURA

[1] bmv2 https://github.com/p4lang/p4c/blob/main/p4include/v1model.p4 (data dostępu 01.01.2023).

[2] DBLATEX. 2017. TRex, https://trex-tgn.cisco.com/trex/doc/trex_book.pdf (data dostępu 01.01.2023).

[3] Intel Deep Insight Network Analytics, https://www.intel.pl/content/www/pl/pl/products/network-io/programmable-ethernet-switch/network-analytics/deep-insight.html (data dostępu 01.01.2023).

[4] Kaczmarek Sylwester, Maciej Krajewski. 2010. „Quality measure of VoIP services in carrier grade systems". ZN WETI PG. Information Technology. Vol. 8, No. 8.

[5] Kaczmarek Sylwester, Maciej Wolff. 2022. „QoS Extensions for Flow-Awareness Networks". Intl Journal of Electronics and Telecommunications. Vol. 68, No. 2: 351-357.

[6] Mandar Joshi. 2021. „Implementation and Evaluation of In-Band Network Telemetry in P4". KTH Royal institute of Technology.

[7] Minicom, https://wiki.emacinc.com/wiki/Getting_Started_With_Minicom (data dostępu 01.01.2023).

[8] Mininet, http://mininet.org/ (data dostępu 01.01.2023).

[9] P4, https://p4.org/ (data dostępu 01.01.2023).

[10] P4 Language Consortium. 2021. "$P4_{16}$ Language Specification". Version 1.2.2, https://p4.org/p4-spec/docs/P4-16-v1.2.2.pdf (data dostępu 01.01.2023).

[11] P4.org. 2017. „P4 Language Tutorial". https://opennetworking.org/wp-content/uploads/2020/12/P4_tutorial_01_basics.gslide.pdf (data dostępu 01.01.2023).

[12] P4.org Applications Working Group. 2020. „In-band Network Telemetry (INT) Dataplane Specification". Version 2.1. https://p4.org/p4-spec/docs/INT_v2_1.pdf (data dostępu 01.01.2023).

[13] QEMU Documentation. 2023. „The QEMU Project Developers". Release 7.2.50. https://readthedocs.org/projects/qemu/downloads/pdf/latest/ (data dostępu 01.01.2023).

[14] Ubuntu 18.04, https://releases.ubuntu.com/18.04/ (data dostępu 01.01.2023).

SKORELOWANIE METRYK QoS DLA USŁUGI PRZEGLĄDANIA STRON INTERNETOWYCH W SIECIACH MOBILNYCH LTE

CORRELATING QoS METRICS FOR WEB BROWSING SERVICE IN LTE MOBILE NETWORKS

Dariusz Zmysłowski[1]; Jan M. Kelner[2]

[1,2] Wojskowa Akademia Techniczna im. Jarosława Dąbrowskiego, Wydział Elektroniki, Instytut Systemów Łączności
ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warszawa, {[1]dariusz.zmyslowski, [2]jan.kelner}@wat.edu.pl

DOI: 10.15199/59.2023.4.9

Streszczenie: **Nowoczesne sieci komórkowe zapewniają dostęp do różnorodnych usług. Z jednej strony operatorzy sieci komórkowych (MNO) są zobowiązani do zapewnienia odpowiedniego zasięgu sieci. Z drugiej strony usługi powinny być świadczone na odpowiednim poziomie. W tym artykule przedstawiamy analizę korelacyjną metryk jakości usług (QoS) i parametrów sygnału radiowego dla scenariusza przeglądania stron internetowych. Niniejsze badanie bazują na kampanii pomiarowej przeprowadzonej w okolicach Warszawy dla dwóch MNO.**

Abstract: **Modern mobile networks provide access to a variety of services. On the one hand, mobile network operators (MNOs) are required to ensure adequate network coverage. On the other hand, the provided services should be performed at an appropriate level. This paper presents a correlation analysis of quality of service (QoS) metrics and radio signal parameters for a web browsing scenario. This research is based on the measurement campaign carried out in the vicinity of Warsaw for two MNOs.**

Słowa kluczowe: **jakość usług, LTE, przeglądanie sieci, RSRP, sieci mobilne, SINR.**

Keywords: **LTE, mobile networks, quality of service (QoS), RSRP, SINR, web browsing.**

1. WSTĘP

Jakość usług w sieciach mobilnych ma obecnie ogromne znaczenie zarówno dla operatorów infrastruktury telekomunikacyjnej MNO (*mobile network operator*), ich użytkowników końcowych (abonentów), jak i regulatorów rynków telekomunikacyjnych.

Dla MNO wiedza z zakresu obiektywnych miar jakości usług QoS (*quality of service*) oraz subiektywnych miar zadowolenia lub irytacji doświadczeń klienta z usługi QoE (*quality of experience*) jest cennym narzędziem do oceny sieci i poziomu świadczonych usług, ale służy również do ustalania przyczyn awarii sieci, niedostępności usługi i wahania wydajności. Ponadto MNO mogą wykorzystać wnioski z analizy statystycznej pomiarów QoS i QoE do oceny wydajności urządzeń i sieci oferowanych przez dostawców na etapie badań pilotażowych PoC (*proof of concept*) nowych rozwiązań sieciowych.

Metryki QoS pozwalają na badanie, analizowanie i ocenę technicznych aspektów funkcjonowania sieci oraz świadczonych usług pod kątem spełnienia wymagań, które zostały pierwotnie postawione. Analizując wartości kluczowych wskaźników wydajności KPI (*key performance indicator*) dla QoE, można scharakteryzować stan usług w danej sieci z perspektywy użytkownika.

Wykorzystanie możliwości oceny parametrów QoS i QoE jest również istotne dla regulatorów rynku telekomunikacyjnego, np. Urzędu Komunikacji Elektronicznej w Polsce. Badania QoS/QoE mogą wskazywać użytkownikom końcowym aktualny stan usług oferowanych przez poszczególne sieci i porównywać je z perspektywy użytkownika. Regulator rynku może również prowadzić badania dotyczące zgodności deklaracji MNO składanych w procedurach aukcyjnych i koncesyjnych rozwoju sieci i usług, ich zakresu, zasięgu i jakości [3,5].

W tym artykule przedstawiono ocenę korelacyjną między parametrami sygnału radiowego a metrykami QoS dla usługi przeglądania stron internetowych (*web browsing*). Taka ocena może być podstawą do modelowania jednego z parametrów na podstawie drugiego [4]. Autorzy [2] proponują innowacyjne podejście do wyboru usług z uwzględnieniem ich korelacji z wymaganiami użytkowników. Z kolei korelacyjna analiza metryk QoS pozwala na wybór nieskorelowanych metryk do oceny MNO w szerszym aspekcie zapewnianej QoS i QoE [5].

2. POMIARY JAKOŚCI USŁUG TELEKOMUNIKACYJNYCH

2.1. Stanowisko pomiarowe

Pomiary QoS zostały wykonane przez firmę Systemics-PAB z wykorzystaniem systemu pomiarowego firmy Rohde & Schwarz. System pomiarowy obejmował SwissQual Diversity Smart Benchmarker Rel. 20.3 w oparciu o terminale pomiarowe z oprogramowaniem SwissQual QualiPoc. W testach, jako końcówki pomiarowe, wykorzystano terminale mobilne Samsung Galaxy S20+ 5G (SM-G986BDS). Terminale te obsługują technologię agregacji nośnych dla wszystkich pasma wykorzystywane przez MNO w Polsce.

Dodatkowo wykorzystano pasywny skaner Rohde & Schwarz TSME6, który służył do pomiaru metryk QoS reprezentujących jakość i moc sygnałów radiowych. Skaner obsługuje wszystkie pasma częstotliwości używane w sieciach komórkowych.

Stanowisko pomiarowe montowane jest w bagażniku tylnym pojazdu, natomiast terminale pomiarowe

umieszczone są w specjalnych obudowach w dachowym boksie na wysokości około 1,8 m [4].

Rys. 1. Pojazdy ze stanowiskiem pomiarowym wykorzystane w badaniach

2.2. Scenariusz pomiarowy

Badania prowadzono w okolicach miejscowości Piaseczno i Góra Kalwaria na południe od Warszawy w dniach od 29 listopada do 14 grudnia 2021 r. W trakcie pomiarów przeprowadzono testy połączeń głosowych i transmisji danych. Testy przesyłania danych obejmowały transmisję na żywo w YouTube, przesyłanie danych oraz przeglądanie stron internetowych (*web browsing*) z wykorzystaniem protokołu HTTP (*Hypertext Transfer Protocol*) [4]. W tym artykule przeanalizowano dane z testu dla scenariusza przeglądania stron internetowych.

Testy przeprowadzono dla dwóch MNO świadczących usługi telekomunikacyjne w Polsce, oznaczonych jako MNO#1 i MNO#2. Ze względów formalnych dane zostały zanonimizowane. W przypadku MNO#1, usługa przeglądania stron realizowana była przy użyciu dwóch technologii, 3G – UMTS (*Universal Mobile Telecommunications Service*) i 4G – LTE (*Long-Term Evolution*). W przypadku MNO#2, wszystkie transmisje zostały wykonane za pomocą LTE.

W przeprowadzonej analizie wykorzystano około 400, 5200 i 6000 zestawów wartości parametrów (tzw. *bins*) odpowiednio dla MNO#1 UMTS, MNO#1 LTE oraz MNO#2 LTE.

3. METRYKI JAKOŚCI USŁUG

W trakcie pomiarów dla każdego punktu pomiarowego (*bin*) określono różne parametry dotyczące MNO, czasu i miejsca pomiaru, zastosowanej technologii (tj. UMTS lub LTE), średniej prędkości pojazdu oraz parametrów QoS. W przypadku usługi transmisji danych czy przeglądania stron internetowych, najbardziej podstawowym parametrem definiującym QoS jest średnia szybkość transmisji Th (*average throughput*), zwana też przepływnością. Oprócz tego w testach rejestrowano inne metryki QoS [4]:

- Th1s – przepływność w pierwszej sekundzie transmisji (*throughput for first second of transmission*),
- TD – czas transmisji (*time duration*),
- SD – czas trwania sesji (*session duration*),
- RTT1 – pierwsza zmierzona wartość czasu transmisji w obie strony (*first round trip time*),
- RTT2 – średni czas transmisji w obie strony dla pozostałej części transmisji (*average round trip time for rest transmission*),
- T1B – czas transmisji pierwszego bajtu (*time to first byte*),
- T50V – czas transmisji 50% zawartości strony (*time to 50% of volume*),
- T500kB – czas transmisji pierwszych 500 kB danych (*time to first 500 kB of transmission*).

Ponadto, w trakcie pomiarów rejestrowano parametry ilustrujące moc i jakość transmitowanych sygnałów radiowych:

- RSRP – moc odbieranego sygnału referencyjnego (*reference signal received power*),
- SINR – stosunek mocy sygnału użytecznego do mocy interferencji i szumu (*signal-to-interference-plus-noise ratio*),
- RSCP – moc kodowa sygnału referencyjnego (*reference signal code power*),
- EC/IO – stosunek mocy nośnej do interferencji w łączu w dół (*downlink carrier-to-interference ratio*).

Parametry RSRP i RSCP określane są w dBm, podczas gdy SINR i EC/IO – w dB. Metryki RSCP i EC/IO wyznaczane są dla sygnałów UMTS, natomiast RSRP oraz SINR – dla LTE. Opis dwóch ostatnich wymienionych metryk oraz ich pomiarów w sieci LTE przedstawiono w [1].

4. ANALIZA KORELACYJNA

4.1. Współczynnik korelacji

Podstawą przeprowadzonej analizy korelacyjnej jest współczynnik korelacji Pearsona PCC (*Pearson correlation coefficient*) definiowany następująco

$$PCC = \rho(X,Y) = \frac{\mathrm{E}\{(X-\mu_X)(Y-\mu_Y)\}}{\sigma_X \sigma_Y} \qquad (1)$$

gdzie $X=\{x_i\}$ i $Y=\{y_i\}$ reprezentują zbiory dwóch analizowanych parametrów, $\mathrm{E}\{\cdot\}$ jest operatorem uśrednienia po zbiorze wartości pozwalającym wyznaczyć wartość oczekiwaną (średnią, przeciętną), μ_X, μ_Y, σ_X oraz σ_Y reprezentują wartości średnie i odchylenia standardowe odpowiednio dla X i Y tj.

$$\mu_X = \mathrm{E}\{X\},\ \sigma_X = \sqrt{\mathrm{E}\{(X-\mu_X)^2\}}, \qquad (2)$$

$$\mu_Y = \mathrm{E}\{Y\},\ \sigma_Y = \sqrt{\mathrm{E}\{(Y-\mu_Y)^2\}}. \qquad (3)$$

PCC służy do oceny korelacji między dwiema metrykami QoS. W tabeli 1 przedstawiono interpretację wartości PCC. W dalszej części artykułu oprócz PCC, wyznaczone zostały również proste regresji liniowej dla wybranych par parametrów QoS.

Warto podkreślić, że ujemna wartość PCC odpowiada przypadkowi, w którym wzrostowi wartości X towarzyszy spadek średnich wartości Y.

Tab. 1. Klasyfikacja korelacji

$\|PCC\|$	Korelacja
0 ÷ 0,3	Brak, bardzo słaba korelacja
0,31 ÷ 0,5	Umiarkowana korelacja
0,51 ÷ 0,7	Silna korelacja
0,71 ÷ 0,9	Bardzo silna korelacja
0,91 ÷ 1	Korelacja pełna

4.2. Korelacja pomiędzy parametrami QoS

Na podstawie zarejestrowanych danych wyznaczono PCC dla każdej pary rozpatrywanych metryk QoS opisanych w Sekcji 3. Uzyskane wyniki przedstawiono w tabelach 2–4 odpowiednio dla MNO#1 UMTS, MNO#1 LTE i MNO#2 LTE.

Tab. 2. PCC pomiędzy metrykami QoS dla MNO#1 UMTS

PCC	TD	RTT1	RTT2	T1B	T50V	T500kB	Th1s	Th	SD	RSCP	EC/IO
EC/IO	-0.18	-0.23	-0.14	-0.38	-0.03	-0.53	-0.10	0.17	-0.21	0.51	1.00
RSCP	-0.16	-0.13	-0.15	-0.26	-0.08	-0.37	0.05	0.16	-0.18	1.00	
SD	0.99	0.02	0.36	0.13	0.81	0.59	-0.18	-0.31	1.00		
Th	-0.29	-0.02	-0.13	-0.06	-0.19	-0.39	0.39	1.00			
Th1s	-0.21	0.23	-0.02	0.00	-0.37	-0.24	1.00				
T500kB	0.56	0.23	0.19	0.57	0.50	1.00					
T50V	0.85	-0.28	0.22	-0.17	1.00						
T1B	0.06	0.46	0.20	1.00							
RTT2	0.31	0.29	1.00								
RTT1	-0.06	1.00									
TD	1.00										
Scale |PCC|	0.00÷0.09	0.10÷0.19	0.20÷0.29	0.30÷0.39	0.40÷0.49	0.50÷0.59	0.60÷0.69	0.70÷0.79	0.80÷0.89	0.90÷0.99	1.00

Tab. 3. PCC pomiędzy metrykami QoS dla MNO#1 LTE

PCC	TD	RTT1	RTT2	T1B	T50V	T500kB	Th1s	Th	SD	RSRP	SINR
SINR	-0.18	-0.12	-0.31	-0.15	-0.26	-0.31	0.26	0.27	-0.19	0.67	1.00
RSRP	-0.15	-0.11	-0.30	-0.12	-0.24	-0.28	0.21	0.23	-0.16	1.00	
SD	1.00	0.12	0.52	0.46	0.81	0.53	-0.48	-0.42	1.00		
Th	-0.40	-0.09	-0.29	-0.49	-0.42	-0.67	0.77	1.00			
Th1s	-0.45	-0.10	-0.26	-0.58	-0.46	-0.57	1.00				
T500kB	0.49	0.31	0.45	0.73	0.67	1.00					
T50V	0.80	0.02	0.58	0.45	1.00						
T1B	0.41	0.40	0.26	1.00							
RTT2	0.50	0.29	1.00								
RTT1	0.06	1.00									
TD	1.00										
Scale |PCC|	0.00÷0.09	0.10÷0.19	0.20÷0.29	0.30÷0.39	0.40÷0.49	0.50÷0.59	0.60÷0.69	0.70÷0.79	0.80÷0.89	0.90÷0.99	1.00

Tab. 4. PCC pomiędzy metrykami QoS dla MNO#2 LTE

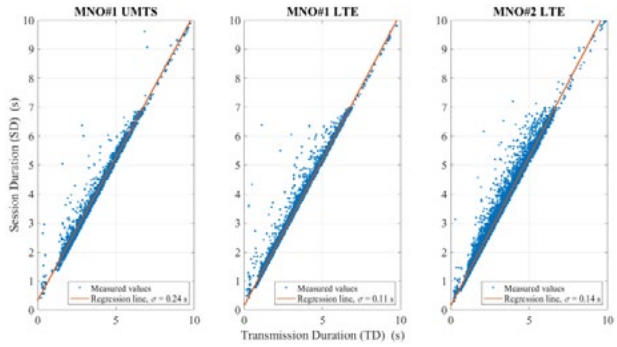
PCC	TD	RTT1	RTT2	T1B	T50V	T500kB	Th1s	Th	SD	RSRP	SINR
SINR	-0.22	-0.15	-0.28	-0.27	-0.33	-0.38	0.25	0.27	-0.24	0.66	1.00
RSRP	-0.18	-0.13	-0.25	-0.20	-0.28	-0.31	0.16	0.19	-0.19	1.00	
SD	0.99	0.24	0.44	0.62	0.85	0.61	-0.51	-0.44	1.00		
Th	-0.40	-0.21	-0.32	-0.60	-0.43	-0.71	0.85	1.00			
Th1s	-0.47	-0.20	-0.29	-0.64	-0.48	-0.65	1.00				
T500kB	0.55	0.41	0.54	0.86	0.77	1.00					
T50V	0.84	0.23	0.52	0.67	1.00						
T1B	0.56	0.43	0.42	1.00							
RTT2	0.42	0.45	1.00								
RTT1	0.18	1.00									
TD	1.00										
Scale |PCC|	0.00÷0.09	0.10÷0.19	0.20÷0.29	0.30÷0.39	0.40÷0.49	0.50÷0.59	0.60÷0.69	0.70÷0.79	0.80÷0.89	0.90÷0.99	1.00

Porównując wyniki zestawione w powyższych tabelach można zauważyć, że są one zbliżone dla wybranej pary metryk QoS, np. PCC pomiędzy Th i TD wynosi odpowiednio –0,29 dla UMTS (MNO#1), –0,4 dla LTE (MNO#1 i MNO#2). Bezwzględna wartość PCC dla dwóch analizowanych parametrów QoS jest zwykle większa dla technologii LTE niż dla UMTS. Ponadto, w przypadku technologii LTE, nieco większe skorelowanie pomiędzy parametrami występuje przeważnie dla MNO#2, np. PCC pomiędzy Th i Th1 wynosi 0,77 dla MNO#1 i 0,85 dla MNO#2, PCC pomiędzy T500kB i T1B wynosi 0,73 i 0,86 odpowiednio dla MNO#1 i MNO#2.

4.3. Regresja liniowa pomiędzy metrykami QoS

Stopień skorelowania pomiędzy dwoma wybranymi parametrami QoS jest również dobrze widoczny, gdy zostaną one zilustrowane na płaszczyźnie X-Y. Jeżeli chwilowe wyniki pomiarów metryk QoS skupiają się wzdłuż pewnej prostej, wówczas występuje silne skorelowanie, tj. wartość bezwzględna PCC powinna być bliska jedności. Mniejsza wartość $|PCC|$ przekłada się na większy rozrzut danych na wykresach.

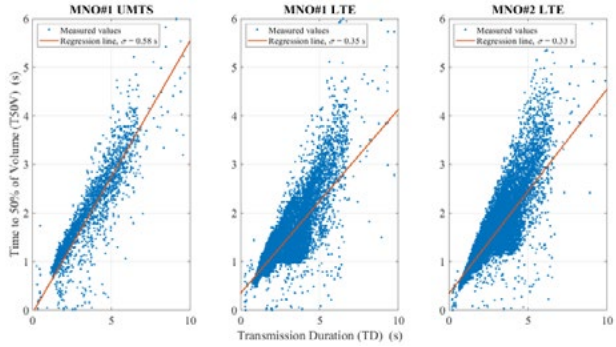
Na rysunkach 2–7 zilustrowano wyniki przykładowych par metryk QoS: SD–TD (rys. 2), T50V–TD (rys.3), Th–Th1s (rys. 4), Th–T500kB (rys. 5). EC/IO–T500kB lub SINR–T500kB odpowiednio dla UMTS i LTE (rys. 6) oraz EC/IO–RSCP dla UMTS lub SINR–RSRP dla LTE (rys. 7).



Rys. 2. SD w funkcji TD dla MNO#1 i MNO#2

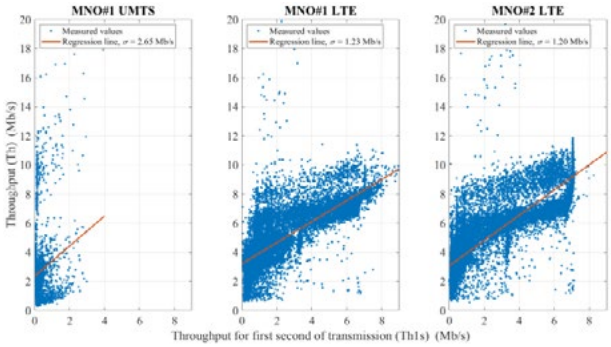
Zilustrowane na powyższych rysunkach rozrzuty wartości pomiarowych dla wybranych par parametrów QoS dobrze odzwierciedlają stopień ich skorelowania. Na przykład, wartości TD i SD przedstawione na rys. 2 są współliniowe, co świadczy do bardzo silnej korelacji. Jest to zgodne z analizą korelacyjną przeprowadzoną w Sekcji 4.2. W tym przypadku PCC wyniósł 0,99 dla UMTS MNO#1 i LTE MNO#2 oraz 1 dla LTE MNO#1.

Dobrym miernikiem korelacji pomiędzy parametrami jest również odchylenie standardowe wyznaczone względem prostej regresji. W tym przypadku wyniosło ono 0,24 s dla UMTS MNO#1, 0,11 s – dla LTE MNO#1 oraz 0,14 s – dla LTE MNO#2.

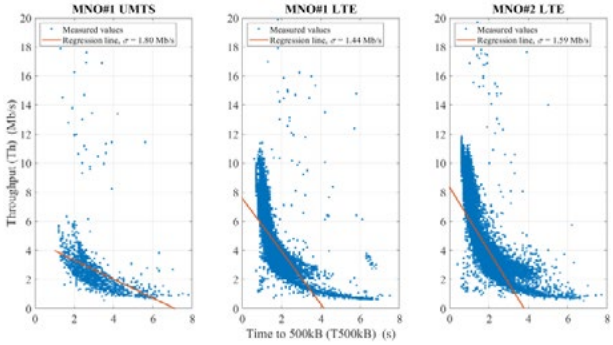


Rys. 3. T50V w funkcji TD dla MNO#1 i MNO#2

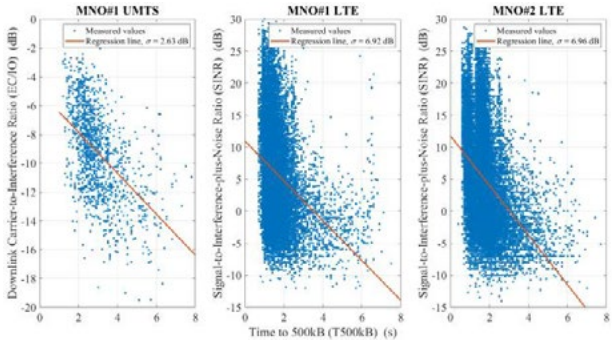
Dla pary T50V–TD zilustrowanej na rys. 3 występuje bardzo silne skorelowanie danych pomiarowych, co pokrywa się z wartościami PCC – 0,85 dla UMTS MNO#1, 0,8 dla LTE MNO#1 oraz 0,81 dla LTE MNO#2. Podobny stopień skorelowania uzyskano dla par metryk jakości sygnałów EC/IO–RSCP dla UMTS lub SINR–RSRP dla LTE (rys. 7). Silne skorelowania występuje także pomiędzy Th i Th1s (rys. 4) dla technologii LTE oraz nieco słabsze – dla UMTS.



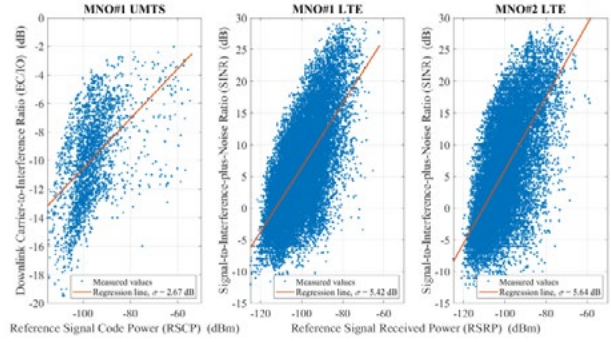
Rys. 4. Th w funkcji Th1s dla MNO#1 i MNO#2



Rys. 5. Th w funkcji T500kB dla MNO#1 i MNO#2



Rys. 6. EC/IO lub SINR w funkcji T500kB dla MNO#1 oraz MNO#2



Rys. 7. EC/IO lub SINR w funkcji RSCP lub RSRP dla MNO#1 i MNO#2

Ujemna wartość PCC dla Th–T500kB (rys. 5) oraz EC/IO–T500kB lub SINR–T500kB (rys. 6) przekłada się na prostą regresji z ujemnym współczynnikiem korelacji.

Uzyskane wyniki pozwalają z jednej strony modelować zmiany metryk QoS na podstawie wartości parametrów sygnału (np. RSRP czy SINR). Z drugiej strony przeprowadzona analiza pozwala na ograniczenie liczby parametrów branych pod uwagę przy ocenie zjawisk lub procesów związanych z QoS.

5. PODSUMOWANIE

Przeprowadzona analiza korelacji danych zarejestrowanych podczas jazdy próbnej wykazała, że wspomniane metryki QoS i parametry sygnału radiowego są ze sobą powiązane. Silną korelację możemy zaobserwować między metrykami czasowymi (np. TD–SD, TD–T50V, T1B–T500kB), przepustowością (tj. Th–Th1s) lub parametrami sygnału (np. SINR–RSRP dla LTE). Pomiędzy niektórymi metrykami QoS a parametrami sygnału radiowego korelacja jest od średniej do słabej. Z jednej strony lepsza jakość kanału i odbieranego sygnału przekłada się na wyższy poziom zapewnianego QoS, co jest oczywistym wnioskiem. Relacje między tymi wskaźnikami nie są jednak trywialne i charakteryzują się pewnymi rozrzutami. Wyznaczenie ich jest możliwe poprzez ocenę regresji liniowej pomiędzy tymi parametrami. W przypadku małych bezwzględnych wartości PCC związek między metrykami jest zwykle nieliniowy. Z drugiej strony wzór Shannona pokazuje, że przepustowość (a pośrednio także przepływność) jest ściśle związana z jakością sygnału (tj. SINR) i szerokością pasma. Dlatego zasoby radiowe udostępniane przez MNO użytkownikom w celu realizacji usługi powinny dodatkowo zostać uwzględnione w analizie korelacji pomiędzy QoS a parametrami sygnału. Przyszłe badania autorów tej publikacji będą koncentrować się na poszukiwaniu nieliniowych zależności między analizowanymi parametrami.

PODZIĘKOWANIA

Autorzy pragną podziękować firmie Systemics-PAB Sp. z o.o. za udostępnienie danych z kampanii pomiarowej, które zostały wykorzystane w niniejszej publikacji.

LITERATURA

[1] Afroz, F., R. Subramanian, R. Heidary, K. Sandrasegaran,, S. Ahmed. 2015. “SINR, RSRP, RSSI and RSRQ measurements in long term evolution networks”. *International Journal of Wireless & Mobile Networks*, 7 (4) : 113–123.

[2] Li D., D. Ye, N. Gao, S. Wang. 2019. “Service selection with QoS correlations in distributed service-based systems”. *IEEE Access*, 7 : 88718–88732.

[3] Mongay Batalla J., S. Sujecki, J. Oko, J.M. Kelner. 2022. “Cost-effective measurements of 5G radio resources allocation for telecom market regulator’s monitoring”. *2022 19th ACM PE-WASUN*, 83–90.

[4] Zmysłowski D., J.M. Kelner. 2022. “Drive test-based correlation assessment of QoS parameters for exemplary measurements scenario in suburban environment”. *2022 18th WEBIST*, 485–494.

[5] Zmysłowski D., J.M. Kelner. 2022. “The concept of application of the Wroclaw taxonomy for QoS assessment in mobile networks”. *2022 18th WEBIST*, 485–494.

QUANTIZATION PARAMETER-BASED VIDEO QUALITY ASSESSMENT IN MPEG CODECS LEVERAGING VMAF

Syed Uddin[1]; Mikolaj Leszczuk[2]; Michal Grega[3]

[1] AGH University of Krakow,, Krakow, Poland, syed.uddin@agh.edu.pl
[2] AGH University of Krakow, Krakow, Poland, mikolaj.leszczuk@agh.edu.pl
[3] AGH University of Krakow, Krakow, Poland, michal.grega@agh.edu.pl

DOI: 10.15199/59.2023.4.10

Abstract: **Video streaming systems are rapidly evolving. There is an increasing demand for high-resolution content such as full hd (FHD) and 4K (UHD). This high-resolution content requires high processing power and storage. Video coding technology is used to compress video content. There are encoding parameters that effect compression efficiency and video quality. It is important to explore those coding parameters which have an impact on video compression and quality. This paper considers MPEG-4, MPEG-AVC, and MPEG-HEVC codecs for testing. The video sequence is encoded using different rate control modes and bitrates. The quality of encoded videos is estimated using the video multimethod assessment fusion (VMAF) objective model. The results show that MPEG-AVC performs significantly better than other codecs. The results also confirm the strong relationship between encoding parameters and video quality. This study provides a foundation for building optimization models in the video quality domain. The study results will be utilized to build robust optimization models for video quality evaluations.**

Keywords: **MPEG, Video codec, VMAF, Video streaming.**

1. INTRODUCTION

Video streaming is growing rapidly, and it is estimated that video traffic will represent 82 % of the total internet consumption. in coming years [1]. As the video demand is increasing, it is challenging to meet user expectations in terms of the quality of experience (QoE) of streaming applications. The high-resolution video such as UHD, 4K and 8K is emerging [2]. These high video resolutions require a high bitrate for transmission which ranges from 5Mbps to 110Mbps [3]. Relatively huge storage and faster processing resources are required to stream high resolution videos.

The video compression technology is used to compress high resolution videos and transmit them efficiently, retaining the quality. Video codecs developed by both MPEG and ITU-T, which compress video, minimize the size of video and also maintain the quality [4]. The MPEG-H264 dominates video compression technology. The H.264 showed promising coding efficiency and reliability for standard definition (SD) The MPEG-HEVC is also a widely known codec. This codec is designed to be valid for H.264 applications and supports high resolution videos [5].

The encoder uses quantization parameters (QP) to determine the data rate. The QP play an important role for encoding video and retain the quality. There should be optimal QP, which ensure the best visual quality and stream video sequence by consuming minimum data rate, which will impact on a network bandwidth [6]. Every codec has its own set of parameter settings. These parameters specify coding pass, rate-control mode, and visual tunning for the codec [7].

Rate control allows video encoder to decide the number of bits spent for a given frame. Rate control is an important step to determine tradeoff between size and quality. There are various rate control modes in practice. These are constant bitrate (CBR), average bitrate (ABR) and constrained encoding (VBV) [8]. The objective video quality model (VQM) is used for the perceptual quality estimation of the videos [9]. There are few metrics used to measure video quality. The video multimethod assessment fusion (VMAF) metric developed by Netflix is a full reference (FR) metric that compares the reference (source) and distorted video sequences to predict the subjective quality [10].

This paper focuses on video quality assessment using bit rate control modes. This paper also studies the VMAF objective quality metric and its behavior when using different quantization parameters by applying various bit rates. The test analysis is carried out with three MPEG codecs.

This paper is organized as follows. Section 2, describes the main work related to the study. Section 3, discusses the research methodology. The final section 4, describes the conclusion and also dedicated to the future work.

2. BACKGROUND

The aim of the video compression is to minimize the bitrate and storage complexity and simultaneously retain video quality. As mentioned before, there is a constant growth and high demand for video streaming services such as video on demand (VOD) and live streaming services. These services are provided by companies such as YouTube, Netflix, and Amazon Video. These service providers must meet the expectations of end users. It is essential to know the requirements of real users and measure their satisfaction and overall quality of experience (QoE).

There is a significant research effort in video codecs compression and the optimization of video sequences. A

study is carried out with modern codecs VVC, SVT-AV1, x265 and VP9. The test sequences were in high definition (HD). This work also utilizes VMAF objective model. The study report significant results on bitrate savings [11]. In this research work HEVC and VVC codecs were considered. The Full high definition (FHD) sequences were used for test. The study reports, that VVC performs better than HEVC in term of compression [12]. A comparison study was carried on VVC, HEVC and AV1 codecs. The author reported that VVC provides the best coding efficiency for higher resolutions. On the other hand, AV1 codec outperformed HEVC [13].

The HEVC, VP9 and AV1 codec compared. The results reveal that AV1 delivered better values as compared to HEVC and VP9. Importantly, among all codecs AV1 takes longer time to encode [14]. A study performed with H.264, H.265 and AV1 codecs. Three models VMAF, SSIM and PSNR calculated. The authors reported that the AV1 has shown the best quality for most bitrates and contents than its competitors [15]. A study concludes that AVC implementation perform better than HEVC, VP9, AVS2 and AV1 in terms of subjective quality using ultra high-definition videos [16]. In a research work HEVC and AV1 codecs evaluates. The study consider ultra high-definition video. Interestingly, the results reveal that both codecs do not differ significantly in most cases [17].

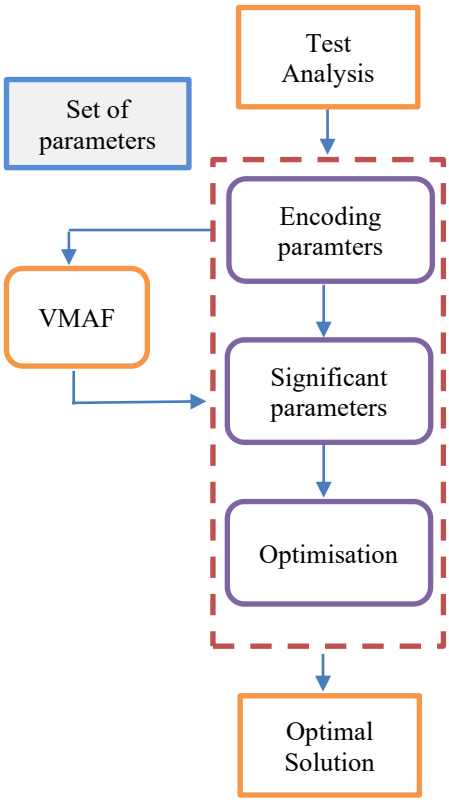


Figure 1 Research methodology flowchart

3. METHODOLOGY

In this section, the testing methodology and procedure are described in detail. The test analysis and results are also presented in this section. The flow chart in Figure 1, demonstrates the research methodology.

3.1. Video Acquisition

The video used for testing is acquired from a test media dataset [18]. The video was in a raw format (YUV) and in Full high definition (FHD) resolution. The FFMPEG tool was used to extract 300 frames from the video. The screenshots are shown in Figure 2:

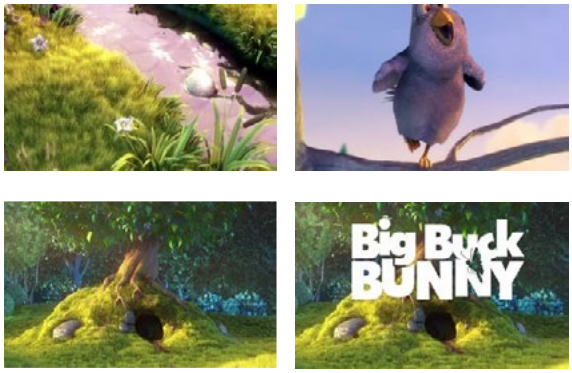


Figure 2 Reference video screenshots

The video sequence was encoded using FFMPEG implementation of codecs MPEG H.264, H.264/AVC, and H.256/HEVC respectively.

3.2. Encoder Settings

Table 1 Video coding parameters

Parameter	*Value*
Sequence duration	12.5 s
Sequence resolution	1920x1080 pixel
Frame Rate	24
Encoder	FFmpeg
Rate control mode	Constant bitrate (CBR)
Compression standards	H.264, AVC, HEVC
Presets	Faster

The table 1, provides a summary of the encoder settings used in testing. Tests were carried out with different combinations of parameters. The video sequence is encoded at constant bit rate (CBR) control mode. The test cases were developed. The target bit rate was set to 1 Mbps, 2 Mbps, 3 Mbps, 4 Mbps, 5 Mbps, and 6 Mbps respectively.

3.3. Video assessment metric

The video multimethod assessment fusion (VMAF) model is deployed in to estimate video quality. The VMAF is highly correlated to subjective scores. The VMAF uses a machine learning algorithm that combines several objective quality features into a single score.

4. RESULTS & ANALYSIS

This section presents the experimental results. The test cases were developed to determine the level of occurrence of Just a Noticeable Difference (JND) [19].

The VMAF model is used to determine JND. The MPEG H.264, AVC and HEVC codecs are used for evaluation. The FFMPEG tool is used for the encoding process. Statistical comparison was made using one-way ANOVA. The VMAF was used as a dependent variable and codec version was used as the factor. The statistical interpretations were done at a confidence level of 95 percent.

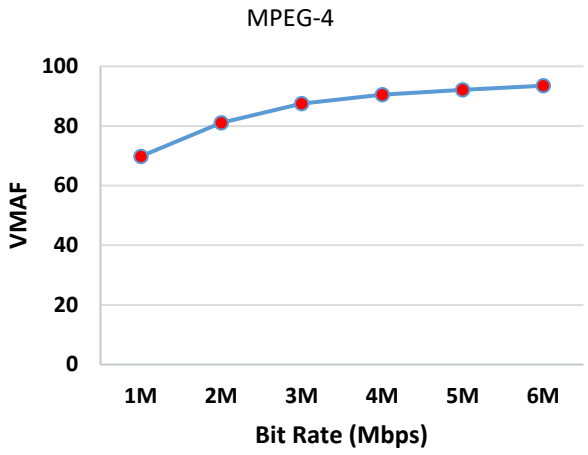


Figure 3. Relationship between bitrate and VMAF for MPEG4

The video sequence encoded by target input bitrate and the output bitrate is obtained as per the input bitrate. The VMAF is calculated for each encoded sequence on a set input bitrate.

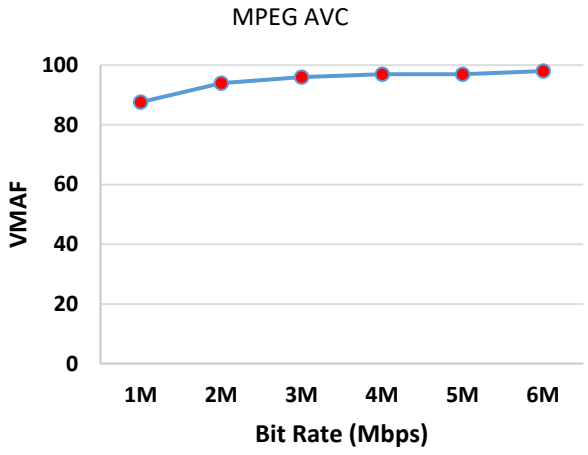


Figure 4. Relationship between bitrate and VMAF for MPEG AVC

The VMAF score of 90 is assumed as an acceptable quality. The VMAF is higher (90) when video sequence was encoded at 6Mbps. The VMAF was lower when the sequence encoded at 1Mbps. A detailed analysis is provided in Figure 3.

Encoding using MPEG-AVC the highest VMAF value was at 6Mbps. When the sequence was encoded using input 4Mbps and 5Mbps the VMAF values shows straight line. The VAMF values does not show significant differences when encoded at bitrates 4Mbps and 5Mbps. The detailed analysis is shown in Figure 4.

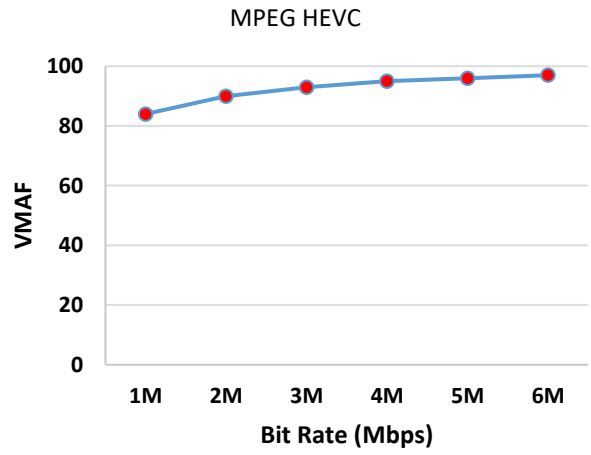


Figure 5. Relationship between input bitrate and VMAF for HEVC

As depicted in Figure 5, the VMAF value increases as bit rate increase. The input bitrate is 1Mbps and it gives result of VMAF (=81). This is the lowest value of quality when sequence encoded at loweset rate. There is slightly difference between quality levels when encoded using 5Mbps and 6Mbps respectively.

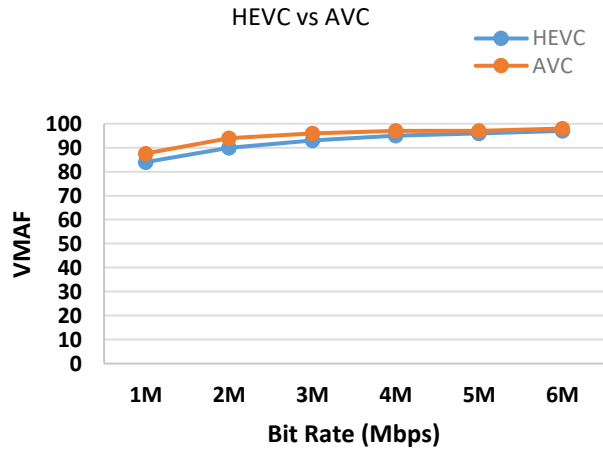


Figure 6. Variability between HEVC and AVC

Interestingly, both HEVC and AVC codecs show different trends when encoded using various rate modes. The trend depicted that AVC is slightly better as compared to HEVC. Thus, AVC outperforms HEVC under given conditions. When encoded at the rate of 5Mbps both HEVC and AVC have similar quality levels which are evident from VMAF scores (+90). The detailed comparison is shown in Figure 6.

The significance is also verified by statistical results. The Mean of VMAF score is compared with codec versions. The Mean of MPEG-AVC (94.93 ± 3.84) was the highest and on the other hand, the Mean value of MPEG-4 (85.75 ± 8.97) was the lowest.

Table 2 Statistical relationship between VMAF and codecs

	Sum of squares	*df*	*Mean square*	*F*	*Significance*
BG	271.6	2	135.8	3.4	.059
WG	593.2	15	39.5		
Total	863.9	17			

Table 2, depicts the statistical relationship between VMAF and codec versions. The (BG) represents between-groups and (WG) represents within-group comparisons. The results demonstrate that there is a significant difference between the video sequence encoded with a specific codec and its relative VMAF scores.

5. CONCLUSION

In this paper video quality is assessed using rate control mode deploying video multimethod assessment fusion (VMAF) objective model. The purpose of this research is to address the video compression and its impact on the quality. There are parameters associated with encoder which potentially impact video compression and computation complexity. The MPEG-4, MPEG-AVC and MPEG-HEVC codecs are used for testing. The video sequence encoded using different input rates. The VMAF scores calculated for the encoded sequence. The optimal results were analyzed and recorded.

The results demonstrate that MPEG-AVC performs better as compared to other considered codecs. The results also prove that encoding parameters have an impact on video quality. The test comparisons have been verified using video quality assessment (VQA). The VMAF scores correlated with encoding bit rates.

This study has limitations. The study conducted using specific video sequence from single source. The findings from this study are not sufficient to generalize. A comprehensive study is required to consider sequences from sources with different spatial and temporal chracterisics.

In the future study, this work will be extended and more parameters should be considered. A robust optimization model based on encoding parameters will be developed. In the next study, a test will be carried out over MPEG-DASH. The encoding time and computational complexity will be considered. More video quality assessment modes should be incorporated.

REFERENCES

[1] CISCO, "Cisco Annual Internet Report (2018-2023) White Paper ," 2020.

[2] S. Ling, Y. Baveye, D. Nandakumar, S. Sethuraman, and P. Le Callet, "Towards Better Quality Assessment of High-Quality Videos," Dec. 2020, pp. 3–9. doi: 10.1145/3423328.3423496.

[3] F. Xie, M. Pourazad, P. Nasiopoulos, and J. Slevinsky, "Determining bitrate requirement for UHD video content delivery," Dec. 2016, pp. 241–242. doi: 10.1109/ICCE.2016.7430597.

[4] A. Punchihewa and D. Bailey, "A Review of Emerging Video Codecs: Challenges and Opportunities," in *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2020, pp. 1–6. doi: 10.1109/IVCNZ51579.2020.9290536.

[5] C. Iza Paredes, "Performance Comparison of H.265/HEVC, H.264/AVC and VP9 Encoders in Video Dissemination over VANETs," 2022.

[6] I. Katsavounidis and L. Guo, "Video codec comparison using the dynamic optimizer framework," in *Applications of Digital Image Processing XLI*, A. G. Tescher, Ed., SPIE, 2018, p. 107520Q. doi: 10.1117/12.2322118.

[7] N. Barman and M. Martini, "H.264/MPEG-AVC, H.265/MPEG-HEVC and VP9 codec comparison for live gaming video streaming," Mar. 2017. doi: 10.1109/QoMEX.2017.7965686.

[8] I. Khan, M. A. Ansari, S. H. Saeed, and K. Khan, "Evaluation and analysis of rate control methods for H.264/AVC and MPEG-4 video codec," *International Journal of Electrical and Computer Engineering*, vol. 8, pp. 2788–2794, Apr. 2018, doi: 10.11591/ijece.v8i5.pp.2788-2794.

[9] Q. Fan, W. Luo, Y. Xia, G. Li, and D. He, "metrics and methods of video quality assessment: a brief review," *Multimed Tools Appl*, vol. 78, Apr. 2019, doi: 10.1007/s11042-017-4848-x.

[10] B. García, L. López-Fernández, F. Gortázar, and M. Gallego, "Practical Evaluation of VMAF Perceptual Video Quality for WebRTC Applications," *Electronics (Basel)*, vol. 8, p. 854, Apr. 2019, doi: 10.3390/electronics8080854.

[11] G. Esakki, A. S. Panayides, V. Jalta, and M. S. Pattichis, "Adaptive Video Encoding for Different Video Codecs," *IEEE Access*, vol. 9, pp. 68720–68736, 2021, doi: 10.1109/ACCESS.2021.3077313.

[12] M. A. Usman *et al.*, "Suitability of VVC and HEVC for Video Telehealth Systems," *Computers, Materials & Continua*, 2021.

[13] A. S. Panayides, M. S. Pattichis, M. Pantziaris, A. G. Constantinides, and C. S. Pattichis, "The Battle of the Video Codecs in the Healthcare Domain - A Comparative Performance Evaluation Study Leveraging VVC and AV1," *IEEE Access*, vol. 8, pp. 11469–11481, 2020, doi: 10.1109/ACCESS.2020.2965325.

[14] T. Uhl, C. Hoppe, and J. H. Klink, "Modern Codecs by Video Streaming under Use DASH Technique: An Objective Comparison Study," in *2020 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 2020, pp. 1–5. doi: 10.23919/SoftCOM50211.2020.9238324.

[15] D. Ashimov, M. G. Martini, and N. Barman, "Quality Assessment of Gaming Videos Compressed via AV1," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1–4. doi: 10.1109/QoMEX48832.2020.9123112.

[16] Z. Li, Z. Duanmu, W. Liu, and Z. Wang, "AVC, HEVC, VP9, AVS2 or AV1? — A Comparative Study of State-of-the-Art Video Encoders on 4K Videos," 2019, pp. 162–173. doi: 10.1007/978-3-030-27202-9_14.

[17] A. V Katsenou, F. Zhang, M. Afonso, and D. R. Bull, "A Subjective Comparison of AV1 and HEVC for Adaptive Video Streaming," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 4145–4149. doi: 10.1109/ICIP.2019.8803523.

[18] Xiph"Xiph.org Test Media." https://media.xiph.org/

[19] W. Wu and B. Song, "Just-noticeable-distortion-based fast coding unit size decision algorithm for high efficiency video coding," *Electron Lett*, vol. 50, pp. 443–444, 2014.

NOWA METODA WYKRYWANIA INTENCJONALNYCH ZAKŁÓCEŃ W TRANSMISJI OFDM

A NEW METHOD OF JAMMING DETECTION IN OFDM TRANSMISSION

Krzysztof Wesołowski

Politechnika Poznańska, Poznań, krzysztof.wesolowski@put.poznan.pl

DOI: 10.15199/59.2023.4.11

Streszczenie: W artykule przedstawiono nową metodę wykrywania intencjonalnych zakłóceń w transmisji OFDM. Polega ona na porównaniu energii w dwóch następujących po sobie odstępach czasowych. Zaproponowano metrykę, na podstawie której podejmowana jest decyzja, czy zakłócenie wystąpiło. Metodę tę zweryfikowano symulacyjnie w modelowanym łączu systemu 5G New Radio (NR), w którym zawarto pełny łańcuch bloków funkcjonalnych warstwy fizycznej łącza. Podane przykładowe wyniki symulacji dotyczą łącza w górę systemu 5G NR.

Abstract: In this paper, we present a new method of jamming detection in OFDM transmission. It is based on the comparison of signal energy in two subsequent time intervals. An appropriate metric is proposed which enables the decision-making to be made if jamming is present. The proposed method has been verified by modeling the 5G New Radio link, which contains a complete chain of functional blocks of the physical layer of the system. The exemplary simulation results show the detection performance in the 5G NR uplink.

Słowa kluczowe: Transmisja OFDM. system 5G NR, intencjonalne zakłócanie.

Keywords: OFDM transmission, 5G NR system, jamming.

1. WSTĘP

Systemy bezprzewodowe piątej generacji (5G) umożliwiają wprowadzenie nowych usług i aplikacji. Jak wiadomo, w ramach systemów 5G rozróżniamy trzy ich główne rodzaje: szerokopasmowy wzbogacony dostęp bezprzewodowy (eMBB - *enhanced Mobile Broadband*), ultraniezawodną komunikację o niskich opóźnieniach (URLLC - Ultra-Reliable Low Latency Communications) oraz masową komunikację między urządzeniami (mMTC - *massive Machine-Type Communications*). W przypadku niektórych usług zapewnienie bezpieczeństwa systemu jest kluczowe dla ich właściwej realizacji. Wśród wielu możliwych ataków na sieci 5G, które mogłyby uniemożliwić realizację wielu usług, intencjonalne zakłócanie sygnału (*jamming*) w warstwie fizycznej jest jednym z podstawowych typów ataków. W niniejszym artykule przedstawimy metodę wykrywania takich zakłóceń w transmisji MIMO OFDM w systemie 5G NR. W wielu publikacjach na temat detekcji takich zakłóceń rozpatruje się niepełny model łącza, co może prowadzić do nie w pełni prawidłowych wniosków dotyczących badanych metod detekcji. W odróżneniu od tych prac w niniejszym artykule przedstawiamy wyniki eksperymentów na pełnym modelu łącza 5G NR zawierającego wszystkie ważne bloki funkcjonalne.

Źródła intencjonalnych zakłóceń mogą funkcjonować według różnych strategii. Najnowszy przegląd sposobów intencjonalnego zakłócania i metod przeciwdziałania mu można znaleźć w [1]. Inteligentne sposoby jammingu w systemach 5G NR opisano w [2]. W pracy [3] urządzenia generujące jamming podzielono na następujące kategorie:

- **Jammer ciągły** emituje sygnał radiowy w sposób ciągły. Sygnał ten reprezentuje losowe bity, które nie są zgodne z żadnym protokołem warstwy MAC.
- **Jammer "zwodniczy"** przesyła częściowo ważne pakiety składające się z ważnego nagłówka i losowego bloku danych.
- **Jammer losowy** zmienia stan pomiędzy uśpieniem a zakłócaniem kanału transmisyjnego. W stanie uśpienia nie generuje on żadnego sygnału, zaś w stania zakłócania działa w jedym z wyżej wymienionych trybów.
- **Jammer reaktywny** stara się zminimalizować energię sygnału zakłócającego przez detekcję użytecznej transmisji a następnie zakłócając ją starając się zakłócić pakiet użytkownika do takiej postaci, aby w docelowym odbiorniku był on odrzucany jako nieważny.

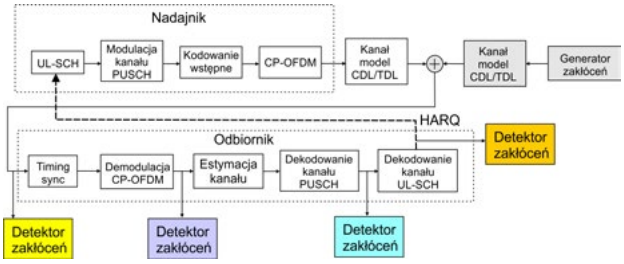
W niniejszej pracy przedstawiamy wyniki dotyczące detekcji działania jammera losowego, który generuje sygnały w formie sygnału szumowego lub w trybie OFDM w łączu w górę systemu 5G.

W literaturze można znaleźć opis szeregu technik detekcji jammingu. Według [4] można je skategoryzować jako techniki wykorzystujące uczenie maszynowe, lub nie stosujące go. W pracy [4] podano przykłady publikacji dla obu kategorii. Jednak z powodu ograniczeń objętości artykułu nie będą cytowane. Metoda przedstawiona w tym artykule nie korzysta z uczenia maszynowego i jest szczególnie prosta w realizacji.

2. MODEL SYSTEMU

Wykonane badania zostały przeprowadzone na przykładzie typowych łączy systemu 5G. W artykule przedstawimy jedynie model łącza w górę zakłócanego przez jammer losowy. Analogiczne wyniki dla łącza w dół zaprezentowano w [5]. Transmisja odbywa się zgodnie z zasadą MIMO OFDM z N_T antenami w stacji ruchomej (terminalu) oraz

N_R antenami w stacji bazowej gNB. Zastosowany został kod kanałowy LDPC zgodnie ze standardami 5G NR oraz procedura HARQ (*Hybrid Automatic Request to Repeat*). Na rys. 1 pokazano schemat blokowy modelowanego systemu.



Rysunek 1: Model transmisji w kanale PUSCH w systemie 5G

Zwróćmy uwagę na fakt, że intencjonalne zakłócenie jest wprowadzane w systemie jako addytywny sygnał dodawany do sygnału odbieranego przez stację bazową z wyjścia kanału transmitującego zakłócenie, którego źródłem jest jammer. Na rys. 1 wskazano potencjalne miejsca detekcji sygnału zakłócającego. Jest to miejsce przed blokiem odbiorczym OFDM (detekcja w dziedzinie czasu), za nim (detekcja w dziedzinie częstotliwości), detekcja na podstawie obserwacji wyjścia dekodera kodu kanałowego oraz na podstawie nietypowego pojawiania się powtórzeń w systemie HARQ. Detekcja w dziedzinie czasu pozwala na wykrycie, czy zakłócenie zostało wprowadzone, jednak nie można łatwo określić jego umiejscowienia na osi częstotliwości. Zaletą detekcji jammingu w dziedzinie częstotliwości jest możliwość identyfikacji zakłóconych bloków zasobów fizycznych (PRB - *Physical Resource Block*). Jednak ceną tej możliwości jest konieczność realizacji równoległych procesów detekcji dla wszystkich bloków PRB, lub co najmniej dla tych, które zajmuje łącze przypisane określonemu użytkownikowi.

3. PROPONOWANA METODA DETEKCJI ZAKŁÓCEŃ

Przeanalizujmy sygnał OFDM otrzymany w dwóch kolejnych przedziałach czasowych o odpowiednio dobranym czasie ich trwania, np. niech będzie to szczelina czasowa systemu 5G NR składająca się z 14 symboli OFDM. Oba przedziały czasowe mogą podlegać ciągłemu przesunięciu w czasie, lub być zsynchronizowane z otrzymanym sygnałem OFDM. Rozpatrzymy drugą z tych możliwości.

Najprostszą metodą detekcji jammingu jest sprawdzanie, czy energia sygnału odbieranego mierzona w określonym oknie czasowym przekracza wartość progową wybraną tak, że w przypadku braku jammingu energia ta jest poniżej progu, zaś w sytuacji występowania zakłócenia – powyżej progu. Niestety wartość tego progu powinna być aktualizowana i dobierana ze względu na aktualną wartość mocy odbieranego sygnału użytecznego, a pośrednio i stosunku mocy synału do mocy szumu, które zależą m.in. od odległości między terminalem a stacją bazową. Zamiast takiego rozwiązania proponujemy zastosowanie następującego kryterium detekcji:

$$d_n = \ln \frac{E_n}{E_{n-1}} \tag{1}$$

w którym E_n oraz E_{n-1} są wartościami energii próbek sygnału odebranego odpowiednio w n-tym i $(n-1)$-szym przedziale na wszystkich antenach w odbiorniku, Tak więc,

$$E_n = \sum_{k=1}^{N_R} \sum_{i=1}^{m} |r_{k,(n-1)m+i}|^2 \tag{2}$$

gdzie k jest indeksem anteny odbiorczej zaś i jest bieżącym numerem próbki otrzymanej w n-tym przedziale czasowym. Załóżmy, że odebrana próbka jest dana za pomocą wzoru:

$$r_{k,j} = \begin{cases} x_{k,j} + \nu_{k,i} + \zeta_{k,j} & \text{jeśli jamming występuje} \\ x_{k,j} + \nu_{k,j} & \text{brak jammingu} \end{cases} \tag{3}$$

zaś $j = (n-1)m + i$. Symbole $x_{k,j}$ oznaczają próbki sygnału użytecznego, $\nu_{k,j}$ są próbkami addytywnego szumu gaussowskiego, natomiast $\zeta_{k,j}$ są próbkami sygnału jammingu. Zauważmy, że jeśli jamming występuje w dwóch kolejnych analizowanych przedziałach czasowych n-tym i $(n-1)$-szym, metryka d_n w (1) jest w przybliżeniu równa zeru zakładając, że statystyki sygnałów nie zmieniają się w tych przedziałach. Natomiast, jeśli jamming pojawi się w n-tym przedziale, zaś nie było go w przedziale poprzednim, metryka d_n jest dodatnia, zaś w sytuacji przeciwnej – ujemna. Zasadniczym zadaniem jest znalezienie progów pomiędzy poszczególnymi przedziałami wartości metryki d_n implikującymi decyzje o wystąpieniu jammingu lub nie. Oznaczmy wartości progowe jako $\pm\Delta$. Wtedy decyzja D_n o detekcji jammingu w przedziale n-tym może być określona na podstawie reguły:

$$D_n = \begin{cases} \text{jeśli } d_n < -\Delta & \text{to } D_n = 0 \\ \text{jeśli } d_n > \Delta & \text{to } D_n = 1 \\ \text{jeśli } -\Delta \le d_n \le \Delta & \text{to } D_n = D_{n-1} \end{cases} \tag{4}$$

gdzie $D_n = 1$ oznacza decyzję o wykryciu jammingu w przedziale n-tym, a $D_n = 0$ oznacza decyzję o braku wystąpienia jammingu w tym przedziale.

Dokładne, teoretyczne wyznaczenie optymalnych progów deycyzjnych $\pm\Delta$ jest zadaniem praktycznie nierozwiązywalnym z powodu niemożności określenia rozkładu prawdopodobieństwa metryki d_n. Tak więc wartości progów zostały wyznaczone na drodze symulacji komputerowych.

Wzór (2) opisuje wielkość energii otrzymanej w dziedzinie czasu. W podobny sposób można wyznaczyć wartości energii w dziedzinie częstotliwości dla każdego z bloków zasobów fizycznych. Umożliwi to określenie wystąpienia jammingu w poszczególnych blokach. Cecha ta jest bardzo przydatna z powodu stosowania w systemie 5G NR zasady wielodostępu OFDMA lub DFT-spread-OFDMA. Tak więc dane łącze może zajmować jedynie część zasobów radiowych rozumianych jako zbiór wykorzystywanych bloków zasobów fizycznych dostępnych w konkretnej szczelinie czasowej. Aby zrealizować detekcje jammingu w poszczególnych blokach, niezbędne jest równoległe działanie procesu jego detekcji w każdym z nich. Proponuje się zastosowanie tej samej miary (1), przy czym nieco odmiennie jest zdefiniowana estymowana energia $E_{n,b}$ w poszczególnych blokach, gdzie b oznacza numer bloku PRB $(b = 1, ..., B)$. Niech liczba bloków PRB w kanale będzie równa B, zaś, jak wiemy, każdy blok wykorzystuje $p = 12$ podnośnych. Energię $E_{n,b}$ zdefiniujemy wtedy jako:

$$E_{n,b} = \sum_{k=1}^{N_R} \sum_{i=1}^{q} \sum_{j=1}^{p} |R_{k,q(n-1)+i,p(b-1)+j}|^2 \tag{5}$$

gdzie R_{k,i_1,j_1} to sygnał odebrany na wyjściu bloku odbiorczego OFDM (FFT) k-tej anteny w i_1-tej chwili $i_1 =$

$(n-1)q+i$, na j_1 podnośnej (j_1-tym wyjściu układu FFT), $j_1 = p(b-1)+j$. W szczelinie jest transmitowanych $q = 14$ symboli OFDM. Reguła decyzyjna (4) pozostaje dalej w mocy, jednak ze wzlgędu na inny zakres wartości sygnałów na wyjściu FFT należy dobrać odpowiednio progi decyzyjne $\pm\Delta$.

4. MODEL SYMULOWANEGO SYSTEMU

Zaproponowana metoda detekcji zakłóceń intencjonalnych w transmisji OFDM została zweryfikowana symulacyjnie za pomocą modelu łącza w górę systemu 5G NR, a konkretnie łańcucha transmisyjnego dla kanału PUSCH (*Physical Uplink Shared Channel*) zastosowanego do realizacji kanału transportowego UL-SCH (*Uplink Shared Channel*). W tym celu zaadaptowano oferowany w pakiecie MATLAB przykład modelowania takiego łącza. Schemat blokowy modelowanego systemu pokazano na rys. 1. Założono, że terminal korzysta z jednej anteny nadawczej ($N_T = 1$) oraz dwóch anten odbiorczych ($N_R = 2$). Szerokość pasma kanału wynosi 10 MHz, zaś odstęp pomiędzy podnośnymi jest równy 15 kHz. Wynika z tego, że w dyspozycj są 52 bloki zasobów fizycznych PRB, każdy o 12-tu podnośnych. W transmisji OFDM zastosowano przedrostek cykliczny o standardowym wymiarze a każda z podnośnych jest modulowana sygnałem QPSK. W kanale UL-SCH zastosowano kod kanałowy LDPC zgodny ze standardem 5G NR. W części odbiorczej słowa kodowe są wyznaczane za pomocą algorytmu iteracyjnego BP (*Belief Propagation*). W celu określenia wielkości bloku transportowego przyjęto sprawność kodowania $R = 193/1024$. Zrealizowana jest również procedura hybrydowego żądania powtórzeń bloków (HARQ) z 16-toma procesami równoległymi. W detekcji sygnału OFDM założono pełną znajomość parametrów kanału transmisyjnego, tak więc uzyskane w symulacji wartości przepustowości należy traktować jako graniczne w stosunku do sytuacji, gdy odbiornik estymuje parametry kanału transmisyjnego.

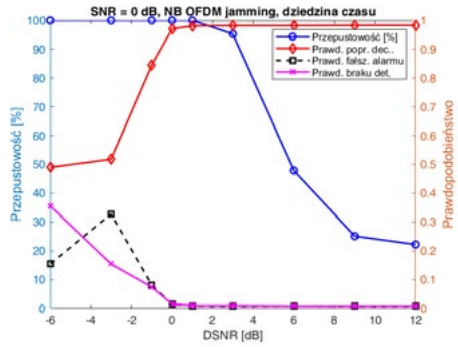
Model kanału transmisyjnego TDL-A (*Tapped Delay Line - A*) jest zgodny ze stadardami modelowania systemów 5G NR i jest oparty na liniach opóźniających MIMO (w naszym przypadku SIMO 1x2). Rozproszenie opóźnień odpowiedzi impulsowej kanału przyjęto równe 30 ns, zaś maksymalna częstotliwość Dopplera wynosi 10 Hz.

Zakłócenie intencjonalne jest modelowane za pomocą źródła szumu lub sygnału OFDM o parametrach podobnych do tych, którymi charakteryzuje się sygnał transmitowany. Sygnał zakłócający jest propagowany przez kanał modelowany również zgodnie z zasadą TDL-A, lecz o mniejszym rozproszeniu opóźnień odpowiedzi impulsowej wynoszącym 20 ns. Sygnał zakłócający może być generowany w całym paśmie lub też w wybranym zakresie częstotliwości o szerokości pasma równej pasmu jednego bloku PRB.

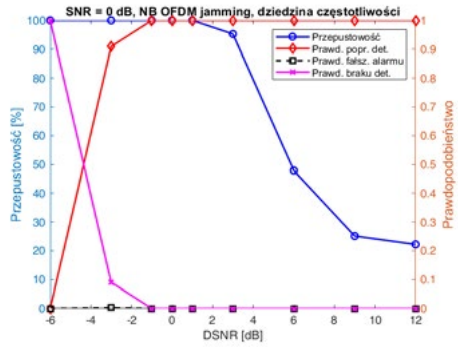
5. WYNIKI SYMULACJI

Jak już wspomniano, zbadano możliwości detekcji jammingu realizowanej zarówno w dziedzinie czasu jak i częstotliwości. Poniżej przedstawimy wybrane wyniki symulacyjne w formie procentowej wartości przepustowości łącza PUSCH względem wartości maksymalnej (wynikającej z zastosowanego kodowania kanałowego LDPC oraz wartościowości modulacji) i prawdopodobieństw detekcji jammingu dla różnych wartości SNR w kanale transmisyjnym oraz różnych wartości poziomu zakłóceń DSNR względem wartości SNR. Poziom mocy sygnału do mocy interferencji imoże więc być wyznaczony z zależności:

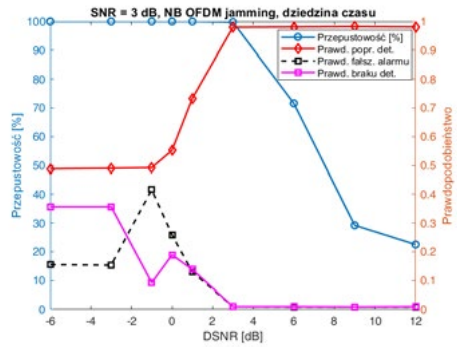
$$SIR_{dB} = SNR_{dB} - DSNR_{dB} \tag{6}$$



Rysunek 2: Przepustowość [%], prawdopodobieństwo poprawnej detekcji jammingu w dziedzinie czasu, fałszywego alarmu i jego niewykrycia - dla SNR = 0 dB w funkcji DSNR [dB]



Rysunek 3: Przepustowość [%], prawdopodobieństwo poprawnej detekcji jammingu w dziedzinie częstotliwości, fałszywego alarmu i jego niewykrycia - dla SNR = 0 dB w funkcji DSNR [dB]

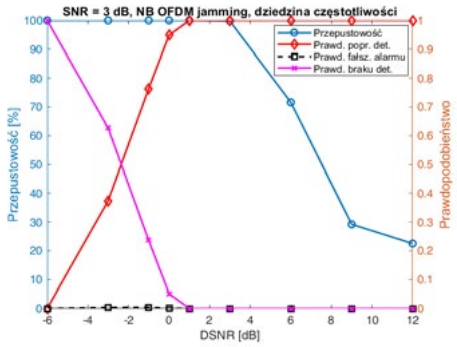


Rysunek 4: Przepustowość [%], prawdopodobieństwo poprawnej detekcji jammingu w dziedzinie czasu, fałszywego alarmu i jego niewykrycia - dla SNR = 3 dB w funkcji DSNR [dB]

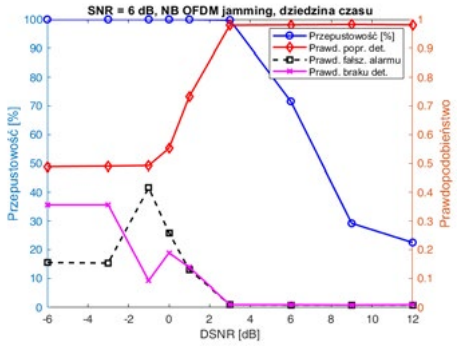
Jamming jest generowany w synchronizacji ze szczelinami czasowymi i założono, że prawdopodobieństwo je-

go wystąpienia wynosi 0.5 (tak więc jest to jamming typu losowego). Oprócz procentowej wielkości przepustowości transmisji, estymowano prawdopodobieństwa poprawnej decyzji dotyczącej detekcji jammingu, prawdopodobieństwo fałszywego alarmu i prawdopodobieństwa braku detekcji jammingu, gdy on rzeczywiście wystąpił.

Na rys. 2 do 7 pokazano jakość działania detektowa jammingu zarówno w dziedzinie czasu jak i częstotliwości. Wybrano przypadek, gdy zakłócany jest jeden z bloków PRB o numerze 10, natomiast SNR wynosi 0, 3 lub 6 dB i jest mierzone w dziedzinie czasu przed demodulatorem. Z rys. 2 do 7 można wywnioskować, że zaproponowany algorytm detekcji działa z dużą niezawodnością. Zwróćmy uwagę, że dla niskich poziomów sygnału jammingu jego detekcja nie jest poprawna, ale dla całości uzyskiwanej jakości transmisji nie ma to znaczenia, ponieważ oddziaływanie jammingu jest dobrze kompensowane przez kodowanie kanałowe i ewentualne procesy powtórzeń HARQ. Gdy poziom jammingu rośnie do tego stopnia, że przepustowość kanału transmisyjnego spada, wykrywalność jammingu zbliża się do 100 procent i jest nieco lepsza w dziedzinie częstotliwości. Prawdopodobieństwa fałszywego alarmu są również bardzo niskie.

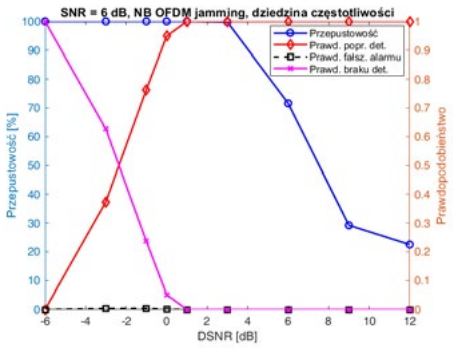


Rysunek 5: Przepustowość [%], prawdopodobieństwo poprawnej detekcji jammingu w dziedzinie częstotliwości, fałszywego alarmu i jego niewykrycia - dla SNR = 3 dB w funkcji DSNR [dB]



Rysunek 6: Przepustowość [%], prawdopodobieństwo poprawnej detekcji jammingu w dziedzinie czasu, fałszywego alarmu i jego niewykrycia - dla SNR = 6 dB w funkcji DSNR [dB]

W ogólnym przypadku proces wykrywania jammingu na wyjściu demodulatora i jego lokalizacja w zakresie pasma konkretnych bloków zasobów fizycznych PRB jest bardziej korzystna niż w dziedzinie czasu. W przypadku zakłóceń wąskopasmowych pozwala ona na lepsze unikanie jammingu przez zmianę przydziału zasobów dla konkretnego łącza.



Rysunek 7: Przepustowość [%], prawdopodobieństwo poprawnej detekcji jammingu w dziedzinie częstotliwości, fałszywego alarmu i jego niewykrycia - dla SNR = 6 dB w funkcji DSNR [dB]

6. WNIOSKI

W artykule zaproponowano metodę wykrywania intencjonalnych zakłóceń (jammingu) w transmisji MIMO OFDM. Weryfikację jej działania pokazano na przykładzie transmisji w kanale transportowym w łączu w górę systemu 5G New Radio. Okazuje się, że metoda ta jest bardzo skutecznym sposobem wykrywania jammingu. Jego detekcja to wstępne zadanie w układzie kompensacji jego wpływu. Obecnie prowadzone są badania mające na celu wykorzystanie jej w tym celu. Metoda została zweryfikowana również w łączu w dół o parametrach transmisji MIMO 8x2, co przedstawiono w pracy [5].

Praca finansowana z projektu CYBERSECIDENT/487845/IV/NCBR/2021.

LITERATURA

[1] H. Pirayesh, and H. Zeng, "Jamming Attacks and Anti-Jamming Strategies in Wireless Networks: A Comprehensive Survey," *IEEE Comm. Surv. Tutorials*, vol. 24 No.2, 2022, pp. 767-809

[2] Y. Arjoune, and S. Faruque, "Smart Jamming Attacks in 5G New Radio: A Review," in Proc. 2020 10th Ann. Computing and Commun. Workshop and Conf. (CCWC), 6-8 Jan. 2020, Las Vegas, USA

[3] A. Hamieh, and J. Ben-Othman, "Detection and Jamming Attacks in Wireless Ad Hoc Networks Using Error Distribution", in Proc. 2009 IEEE Int. Conf, Commun. (ICC), 14-18 June 2009, Dresden, Germany pp.1-6

[4] Y. Arjoune, F. Salahdine, Md. Sh. Islam, E. Ghribi, and N. Kaabouch, "A Novel Jamming Attacks Detection Approach Based on Machine Learning for Wireless Communication", in Proc. 2020 Int. Conf. on Information Networking (ICOIN), 7-10 Jan. 2020, Barcelona, Spain, pp.1-6

[5] K. Wesołowski, "A Simple Algorithm for Jamming Detection in OFDM Systems", IEEE Vehicular Technology Conference, 2023 Spring, Florencja, 20-23 czerwiec, 2023.

DETEKCJA POJAZDÓW W SYSTEMACH 6G Z MODULACJĄ OTFS

VEHICLE DETECTION IN 6G SYSTEMS WITH OTFS MODULATION

Pavel Karpovich [1,2]; Tomasz P. Zielinski[2];

[1] Institute of Telecommunications AGH, Krakow karpovic@agh.edu.pl,tzielin@agh.edu.pl
[2] Nokia Solutions and Networks, Krakow, pavel.karpovich@nokia.com

DOI: 10.15199/59.2023.4.12

Streszczenie: Zaproponowana ostatnio modulacja Orthogonal Time Frequency Space (OTFS) jest bardziej odporna na wąskopasmowy efekt Dopplera, niż technika zwielokrotniania z ortogonalnym podziałem częstotliwości (OFDM - Orthogonal Frequency-Division Multiplexing), używana w standardzie 5G. W referacie zaprezentowano jak modulacja OTFS z dodanym sygnałem pilota, w formie losowego przyrostka, może być z sukcesem użyta w standardzie 6G do jednoczesnej transmisji danych oraz do detekcji szybko poruszających się pojazdów. Porównano dwa algorytmy detekcji obiektów, które są możliwe do zastosowania w transmisji RP-OTFS.

Abstract: The recently introduced orthogonal time frequency space modulation (OTFSM) is more robust to large narrow-band Doppler frequency shift than the orthogonal frequency division multiplexing (OFDM), used in the 5G standard. In this paper it is shown how the telecommunication OTFSM-based signal with random padding can be used with success in the 6G standard for detection of high-speed vehicles. Two approaches for detecting targets during the random padded OTFS based transmission are compared in the paper.

Słowa kluczowe: 5G, 6G, OFDM, OTFSM, radar.

Keywords: 5G, 6G, OFDM, OTFSM, radar.

1. INTRODUCTION

In last few years, the scientific community attention has been focused on the discussion of next generation 6G communication. There are a lot of publications about what applications will drive the 6G network and what technologies should be included in the 6G standard to satisfy their requirements [1] [2]. Among large number of proposals, there are some that are most common, such as a terahertz wave and an integrated sensing and communication (ISAC) [3] [4]. This paper addresses a problem of adding a radar functionality to the communication systems of the future which will use higher frequency carriers and support high-mobility users.

The usage of terahertz band is challenging. Even relatively slow objects could generate very high Doppler frequency shifts. The strong Doppler effect limits the usage of the orthogonal frequency division multiplexing (OFDM) waveform which is at present de-facto a standard waveform in telecommunication systems (e.g. DVB-T2, Wi-Fi, LTE, 5G [5]). The OFDM is based on assumptions that linear convolution of the signal and the channel impulse response can be replaced by circular convolution, and that the channel impulse response is time-invariant or almost time-invariant. This allows to do a very fast and simple channel impulse response estimation. In case of the strong Doppler environment the assumption about constant channel impulse response is no longer valid since any channel coefficient can rotate in complex plane all the time due to the Doppler effect. Using OFDM in such conditions leads to errors in channel estimation and equalization, and eventually to inter-carrier-interference (ICI) and subsequently errors in bit detection.

Increasing sub-carrier spacing (SCS) in OFDM helps to deal with the strong Doppler frequency shift. However, this operation will increase also the OFDM cyclic prefix overhead and reduce transmission efficiency [5]. In order to eliminate the mentioned above disadvantage of the OFDM, the orthogonal time frequency and space (OTFS) modulation was recently introduced in [6]. Due to its unique features it is seriously treated as one of possible 6G waveforms [7].

In this article simulation results for an ISAC system using the OTFS waveform are shown. We will start with the OTFS waveform description, present the delay-Doppler domain used in OTFS and discuss different pilot configurations exploited in it. Next, we will introduce the ISAC system using the OTFS waveform. Finally, in experimental part, we will show results from simulation of a radar part of the discussed RP-OTFS-based ISAC system.

In work [8] results from simulation of the communication part of the RP-OTFS transmission system were presented while this paper addresses simulation of the radar part of the system only. Practical verification of the general RP-OTFS based transmission and sensing concept was already presented in [9].

2. ORTHOGONAL TIME FREQUENCY AND SPACE

The concept of the OTFS is shown in the figure 1 [8] [9]. In comparison to OFDM, the OTFS is a two-dimensional modulation technique. In case of OTFS the modulation process looks as follows. At the beginning modulated IQ/QAM symbols are put into elements of the matrix **A** in figure 1, i.e. on the grid in a delay-Doppler (DD) do-

main. Then, the inverse Zak transform (inverse Fourier transform over the Doppler axis) [10] is used to transform (demodulate) data from the DD to a fast time - slow time (TT) domain. Finally, the obtained samples are reshaped from a matrix into a vector. The DD grid usage for data modulation makes the OTFS waveform attractive for ISAC since it is "native" domain for radars.

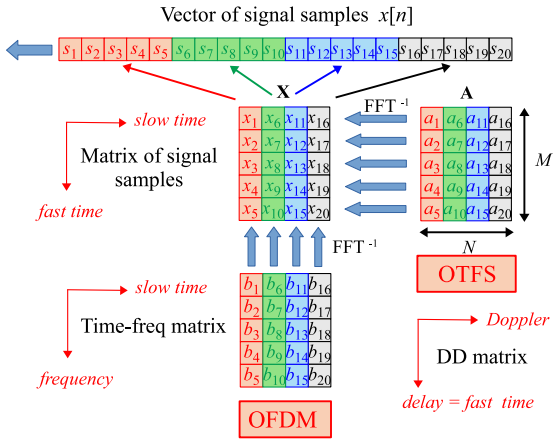


Figure 1: The OTFS concept

2.1 The delay-Doppler grid

Names of the DD grid directions reflect their physical sense. The delay direction (the first D in DD) consists of adjacent samples from time domain, the Δt between samples is small. This direction is suitable for detecting small time changes in a observed signal. For example, in multi-path propagation environment, the difference between paths is not very big and can be estimated in the delay direction of the DD grid. But the delay direction is not suitable for observation of long time processes like the Doppler effect because Doppler frequencies are usually very small and require more time for estimation. In the Doppler direction (the second D of DD) only every Mth sample from time domain is used and this allows to estimate long time signal changes using FFTs with small sizes.

Parameters of the DD grid should be chosen taking into account that the sent OTFS modulated waveform will be used, both, for digital data transmission and moving vehicles detection. As sources of multi-path reflections could be treated as "radar targets", the OTFS DD grid should fulfill, both, telecommunication and radar requirements. The DD grid has two parameters: M — the number of samples in the delay (fast time) direction, and N — the number of samples in the Doppler direction. Looking at figure 1, we can say that in Doppler direction a signal is practically decimated by M. Hence taking into account the Nyquist theorem, the maximum Doppler offset that can be estimated using such DD grid is $f_{dmax} = \pm\frac{f_s}{2M}$, where f_s is a sampling rate. Resolution in the Doppler direction depends on N: increasing N and keeping f_s and M constant will increase the FFT length and the Doppler resolution. The resolution in delay direction depends only on f_s. Choosing f_s, M, N and carrier frequency f_c one can optimize the OTFS-based radar and digital transmission.

For example, lets choose the DD grid parameters for radar detection of many moving cars (reflections from stationary objects are not interesting for us). For maximum car speed of 60 m/s, carrier frequency $f_c = 52.6$ GHz (the maximum carrier frequency for 5G FR2), sampling ratio $f_s = 50$ MHz and maximum $M = 1190$, we can assume that the maximum Doppler frequency shift is equal to 21 kHz. Then, by fixing $M = 1024$ and changing N, one can get different resolution of velocity estimation, changing from 9 m/s (for $N = 8$) to 0.1 m/s ($N = 512$), where values of $N = 8$ and $N = 512$ are exemplary ones.

2.2 Pilots configurations

As OFDM, the OTFS uses pilots for estimation of a channel impulse response (CIR). Their configurations are different. Here we will discuss two types of pilot placement strategies, shown in figure 2: a zero-padded one (ZP-OTFS) and a random-padded one (RP-OTFS). In both configurations the DD matrix **A** is divided into two parts: the data zone and the pilots zone. Every carrier in the DD grid is assigned to the pilot or data zone only, not to both of them the same time.

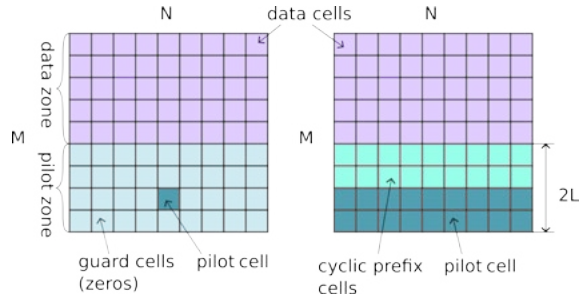


Figure 2: Delay-Doppler grid of the ZP-OTFS (left) and RP-OTFS (right)

2.2.1 ZP-OTFS

In the ZP-OTFS, the pilot has a form of a rectangular zone of the DD matrix **A**, shown in figure 1, which is filled with zeros and have only one non-zero carrier in its center. We will call this non-zero carrier a pilot pulse. In case of the ZP-OTFS, the length of the pilot zone in the delay direction is twice bigger than length of the channel impulse response. In the Doppler direction the pilot zone usually makes use of all cells, as shown in figure 2.

Due to zeros surrounding the pilot pulse, the channel estimation process becomes very simple in the ZP-OTFS. There is also no interference between pilot and data zones as well as no ZP-OTFS symbol interference. In case of ZP-OTFS the channel impulse response is estimated by division of every cell of the received pilot zone by the known, transmitted pilot pulse (some threshold should be introduced here in order not to neglect reflection free samples of the pilot zone). The main disadvantage of the ZP-OTFS is low energy efficiency, because the pilot zone is very sparse.

2.2.2 RP-OTFS

The recently introduced RP-OTFS [8] [9] is designed to correct deficiencies of the ZP-OTFS. Here the pilot zone is filled by short OFDM symbols, treated as pilots, with random data inside — see figure 2. In case of the ZP-OTFS, discussed earlier, the data zone is generated in the DD domain and transformed to fast-slow time domain by the inverse Zak transform. In turn, in case of the RP-OTFS,

OFDM symbols of pilots are inserted directly into fast-slow time grid (without the inverse Zak transformation). Absence of zeros in the pilot zone increases signal-to-noise ratio (SNR) and causes that the RP-OTFS application is more efficient than the ZP-OTFS in CIR estimation what is very important for both for communication and radar.

The CIR estimation begins with conventional OFDM channel estimation with the only difference that we treat the whole OFDM symbol as a pilot. After that, when all CIR momentum estimates are found using all OFDM symbols (having transmitted and received pilots one can easily estimates CIR taps from them), we transform the matrix of CIR taps to the DD domain by the Zak transform, i.e. by performing FFT over the CIR matrix rows. Note, that in the RP-OTFS the Zak transform is performed upon CIR estimates, do not upon time samples of OFDM symbols which were used for CIR calculation.

There are two disadvantages of the RP-OTFS application. Firstly, the length of cyclic prefix (CP) of the OFDM-based pilot should be equal to the OFDM symbol length, i.e. it is long and the CP overhead reduces the achievable bit-rate. Secondly, we assume that the CIR is quasi time-invariant and, therefore, we can not use long OFDM pilots for very high frequency Doppler channels.

3. INTEGRATED SENSING AND COMMUNICATION (ISAC)

In case of ISAC [3][4], usually, the communication processing is the same as in conventional system. In this paper we are concentrating our attention on peculiarity of RP-OTFS radar processing since efficiency of the RP-OTFS based communication sub-system has been already tested [8] [9]. Two approaches of target detection are analyzed: correlation-based and pilot-based.

The first correlation-based method origins from classical radar processing in which a cross ambiguity function is used [11]: transmitted, reference signal (known, re-modulated in the receiver or acquired by special reference antenna) is shifted in time and frequency and correlated with the received, surveillance signal. The problem of the correlation base radar approach is that usually it is hard to find weak signal reflections, coming from small, moving objects, on the background of strong signal reflections caused by buildings (the radar clutter problem) [12].

In the second pilot-based approach of vehicle detection transmitted pilots, known in the receiver, are used to CIR estimation [13]. In case of reflections coming from moving vehicles some CIR taps are complex-value numbers that oscillates in time with frequency of Doppler frequency shift caused the reflecting object movement. Here, we treat radar targets as sources of multi-path propagation. By CIR analysis we can retrieve information about signal reflections and about reflecting objects.

The pilot based ISAC system requires non distorted CIR estimates for Doppler frequency shifts extraction. As mentioned in the introduction, high Doppler objects can not be detected by OFDM. This also limits application of pilot-based radars making use of OFDM-based pilots.

4. EXPERIMENTAL PART

In experimental part we simulated a radar performance of the discussed RP-OTFS-based ISAC system. Parameters of the applied OTFS-based signal was following: size of the grid in delay and doppler direction 64x256 (MxN), length of the pilot zone 16 (meaning of L is explained in fig. 2), modulatiotion 4-QAM, carrier frequency 4 GHz and bandwidth 20 MHz. In simulation we used different target velocities in order to test the system performance in different conditions.

Delay-Doppler (distance-velocity) radar maps for a target moving with velocity about 139 m/s (500 km/h), calculated for both tested radar approaches (correlation based and pilot based ones), are shown in figure 3. In both methods integration/observation time 100 milliseconds was used. Input signal had signal-to-noise ratio (SNR) equal to 0 decibels. In both cases one can clearly see sharp peaks in the delay-Doppler (distance-velocity) matrix which correspond to parameters of moving vehicles. However for CAF two additional lower peaks are visible which are generated by the CP of the pilot part of the RP-OTFS waveform. As in case of the pilot-based approach we eliminate CP from signal processing chain, such peaks are missing in DD map of this method. In case of correlation-based radar mean level of background side-lobes, surrounding the detection peak, is equal to about -30 decibels while for pilot-based radar -40 decibels.

In figures 4 and 5 processing gain charts for both discussed RP-OTFS-based radars are shown, i.e. expressed in decibels root mean square (RMS) value of the method noise floor (visible in figure 3) as a function of signal to noise ratio (SNR) of an input signal. Simulated maximum vehicle speed (v_m) was equal to 50 (13.9 m/s) albeit 500 km/h (139 m/s) and integration/observation time (T_i) was varying from 10 ms to 200 ms. In figure 6 both tested RP-OTFS-based radars are compared: it is seen that the pilot-based version outperforms the correlation-based one in DD detection hight, i.e. in noise robustness.

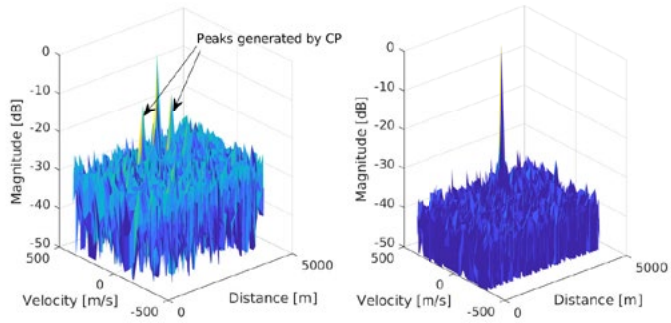


Figure 3: DD map for correlation-based radar (left) and pilot based radar (right)

5. DISCUSSION

The main limitation factor in case of the correlation based RP-OTFS radar is its high level of CAF side-lobes, resulting in significantly lower output SNR in comparison to the pilot-based radar. Figures 4 and 5 confirm quantitative conclusions which can be drown from figure 3.

As mentioned before, in the development of the discussed RP-OTFS-based ISAC system we have assumed that channel pulse response is quasi time-invariant in the pilot zone. In case of high-mobility Doppler channels this assumption is fulfilled only approximately. This fact will limit the maximum processing gain of the presented pilot based RP-OFTS radar. Consequences of this method

drawback will increase for higher velocities as it is visible in fig. 5. The same effect will be observed also when the pilot zone length will be increased. Nevertheless, obtained results confirm that the pilot-based radar outperforms the correlation-based one in terms of noise robustness.

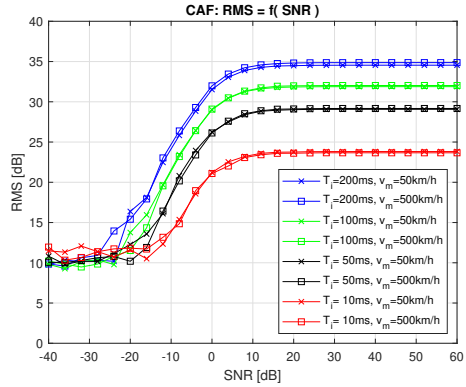


Figure 4: SNR curves for correlation-based radar

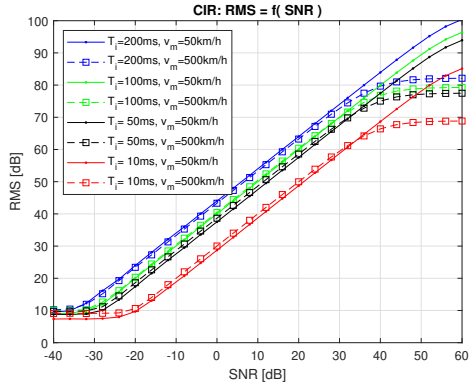


Figure 5: SNR curves for pilot-based radar

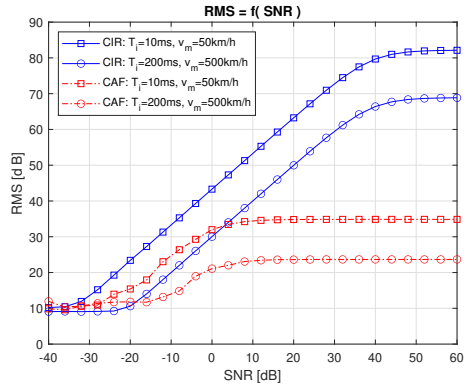


Figure 6: Comparison of SNR curves for correlation and pilot-based radars

6. CONCLUSION

Two moving vehicles detection approaches based on the RP-OTFS ISAC system were compared in this paper. The main limitation factor of the correlation based radar method is high level of CAF side-lobes, apart from existence of two additional peaks in CAF which are caused by repetition of the pilot samples. Detection of targets with low radar cross section on the background of strong background signal, so called clutter, e.g. direct path signal, is very challenging here. Presence of many ghost peaks in the delay-Doppler (distance-velocity) map makes subsequent processing steps in this method very challenging.

In turn, the pilot based RP-OTFS radar is characterized by lower level of side-lobes in the delay-Doppler map and it does not have extra peaks caused by the repeating pilot samples. But this approach is sensitive to the quality of the channel impulse response estimation. In order to minimize error of the channel impulse response estimate, and in consequence error of the moving object detection, we need to keep pilot zone as short as possible.

REFERENCES

[1] H. Tataria et al., "6G Wireless Systems: Vision, Requirements, Challenges, Insights, Opportunities," Proc. IEEE, vol. 109, no. 7, pp. 1166-1199, July 2021.

[2] W. Saad, M. Bennis and M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems," IEEE Network, vol. 34, no. 3, pp. 134-142, May/June 2020.

[3] F. Liu et al., "Integrated Sensing and Communications: Toward Dual-Functional Wireless Networks for 6G and Beyond," IEEE J. on Selected Areas in Comm., vol. 40, no. 6, pp. 1728-1767, June 2022.

[4] Z. Wei et al., "Integrated Sensing and Communication Signals Towards 5G-A and 6G: A Survey," IEEE Internet of Things Journal, early access, 2023.

[5] Josue Flores de Valgas, Jose F. Monserrat, Hüseyin Arslan, "Flexible Numerology in 5G NR: Interference Quantification and Proper Selection Depending on the Scenario", Mobile Information Systems, vol. 2021, Article ID 6651326, 9 pages, 2021.

[6] R. Hadani et al., "Orthogonal Time Frequency Space Modulation," 2017 IEEE Wireless Comm. and Networking Conf. (WCNC), San Francisco, CA, USA, 2017, pp. 1-6, 2017.

[7] Z. Wei at al., "Orthogonal Time-Frequency Space Modulation: A Promising Next-Generation Waveform," IEEE Wireless Comm., vol. 28, iss. 4, pp. 136-144, 2021.

[8] P. Karpovich and T. P. Zielinski, "Random-Padded OTFS Modulation for Joint Communication and Radar/Sensing Systems," 2022 23rd Int. Radar Symp. (IRS), pp. 104-109, Gdansk 2022.

[9] P. Karpovich et al., "Field Tests of a Random-Padded OTFSM Waveform in a Joint Sensing and Communication System," IEEE ICC Int. Communications Conf., Rome 2023.

[10] H. Bolcskei and F. Hlawatsch, "Discrete Zak transforms, polyphase transforms, and applications," in IEEE Trans. on Signal Processing, vol. 45, no. 4, pp. 851-866, April 1997.

[11] M.A. Richards, "Fundamentals of Radar Signal Processing," McGraw-Hill Education, 2014.

[12] P. Karpovich et al., "Practical Results of Drone Detection by Passive Coherent DVB-T2 Radar," 21st Int. Radar Symp. (IRS), pp. 77-81, Warsaw 2020.

[13] M. Braun et al., "Parametrization of joint OFDM-based radar and comm. systems for vehicular applications," 2009 IEEE Int. Symp. on Personal, Indoor & Mobile Radio Comm., pp. 3020-3024, Tokyo 2009.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

Badanie Sieci Massive MIMO o Architekturze Zorientowanej na Użytkownika

Evaluation of User-Centric Massive MIMO Network

Marcin Hoffmann[1];

[1] Politechnika Poznańska, Poznań, marcin.hoffmann@put.poznan.pl

DOI: 10.15199/59.2023.4.13

Streszczenie: W nowoczesnych sieciach mobilnych 6G, wykorzystujących technikę M-MIMO (ang. Massive Multiple-Input-Multiple-Output) prawdopodobne jest zastosowanie architektury zorientowanej na użytkownika, w której każdy użytkownik jest obsługiwany jednocześnie przez wszystkie stacje bazowe. W tej pracy taki system został zbadany z użyciem zaawansowanego symulatora, wykorzystującego model kanału z tzw. trójwymiarowym śledzeniem promieni. Otrzymane wyniki pokazują, że architektura zorientowana na użytkownika zwiększa przepływności osiągane przez użytkowników charakteryzujących się najsłabszymi warunkami radiowymi ponad trzykrotnie.[1]

Abstract: The future 6G networks are expected to utilize large antenna arrays and follow the user-centric architecture, where the user is being served by all base stations. This work evaluates such a system within an advanced system-level simulator, which utilizes an accurate 3D Ray-Tracing radio channel model. Results show that the novel user-centric network architecture can increase the cell-edge users' throughput by a fold of 3.

Słowa kluczowe: 6G, Massive MIMO, sieć zorientowana na użytkownika, trójwymiarowe śledzenie promieni

Keywords: 6G, Massive MIMO, User-Centric networks, 3D Ray-Tracing

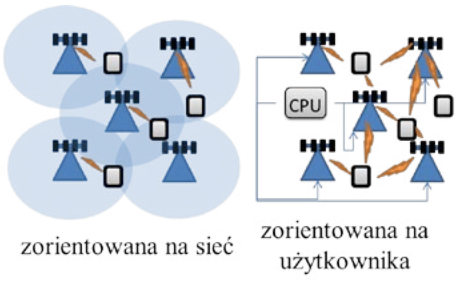
1. WSTĘP

Żeby sprostać rosnącym wymaganiom użytkowników sieci mobilnych konieczne jest wykorzystanie technik transmisyjnych umożliwiających zwiększenie efektywności widmowej. Dla obecnie wdrażanych sieci 5G i przyszłych sieci 6G taką techniką jest tzw. M-MIMO (*ang. Massive Multiple-Input-Multiple-Output*) [1]. Zakłada ona wykorzystanie wieloelementowych macierzy antenowych, które pozwalają na tworzenie wąskich wiązek sygnału radiowego skierowanych do poszczególnych użytkowników. Dzięki temu możliwa jest multipleksacja przestrzenna użytkowników, którym przydzielone są te same zasoby czasowo-częstotliwościowe. Systemy 5G wykorzystujące M-MIMO typowo zakładają obsługę jednego użytkownika przez jedną stacje bazową. Jest to tzw. architektura zorientowana na sieć. W takiej architekturze pokrycie sygnałem radiowym jest nierównomierne: użytkownicy znajdujący się blisko stacji bazowej odbierają znacznie silniejszy poziom sygnału, od tych, którzy znajdują się dalej. W związku z nierównomiernym pokryciem komórek sygnałem radiowym, istnieje silna dysproporcja pomiędzy przepływnościami oferowanymi użytkownikom znajdującym się blisko stacji bazowej, a tymi będącymi na skraju komórki. Z tego względu dla systemów 6G rozważana jest architektura zorientowana na użytkownika, połączona z techniką M-MIMO (*ang. User-Centric-Cell-Free M-MIMO*) [9]. W sieci M-MIMO zorientowanej na użytkownika znika koncepcja komórek, a użytkownik może być obsługiwany przez wszystkie otaczającego go stacje bazowe. W ten sposób użytkownik znajduje się efektywnie w środku wirtualnej komórki, co pozwala znacznie zmniejszyć dysproporcje pomiędzy przepływnościami różnych użytkowników. Wiele prac, w których autorzy badają możliwości systemów o architekturze zorientowanej na użytkownika zakłada bardzo proste modele systemu [7], tzn. system jest wąskopasmowy, podczas gdy systemy 5G powszechnie wykorzystują modulacje wielotonowe i wielodostęp z ortogonalnym podziałem częstotliwości OFDMA (*ang. Orthogonal Frequency-Division Multiple Access*). Ponadto stosowane są proste, statystyczne modele kanałów radiowych często nieuwzględniające istotnych z punktu widzenia M-MIMO korelacji przestrzennych między użytkownikami. Istnieje więc zapotrzebowanie na zbadanie systemów M-MIMO o architekturze zorientowanej na użytkownika w środowisku możliwie bliskim rzeczywistej sieci, np. uwzględniających wykorzystanie techniki OFDMA oraz dokładny model kanału radiowego bazujący na tzw. trójwymiarowym śledzeniu promieni (*ang. 3D Ray-Tracing*). Ponadto należy rozważyć wielowarstwowość przyszłych systemów 6G składających się z różnych bloków funkcyjnych odpowiedzialnych m.in. za przydział zasobów czasowo częstotliwościowych, prekodowanie i dobór schematów modulacji i kodowania. Celem tej pracy jest zbadanie zysków wynikających z zastosowania architektury sieci M-MIMO zorientowanej na użytkownika w porównaniu do tradycyjnej architektury zorientowanej na sieć za pomocą zaawansowanego symulatora komputerowego, opracowanego przez autora. W dalszych częściach tej pracy, w Rozdziale 2 przedstawiona zostanie koncepcja sieci M-MIMO o architekturze zorientowanej na użytkownika. W Rozdziale 3 omówione będą wyzwania związane z praktyczną implementacją takiej architektury sieci. Rozdział 4 zawiera opis zaproponowanego zaawansowanego środowiska symulacyjnego, służącego do badania sieci M-MIMO zorientowanej na użytkownika. Wyniki symulacji przedstawione są w Rozdziale 5, a wnioski są sformułowane w Rozdziale 6.

[1]Praca powstała w ramach projektu PRELUDIUM finansowanego przez Narodowe Centrum Nauki nr 2022/45/N/ST7/01930.

2. SIEĆ ZORIENTOWANA NA UŻYTKOWNIKA

Na Rys. 1 architektura systemu M-MIMO zorientowanego na sieć została porównana z architekturą zorientowaną na użytkownika. W pierwszym przypadku użytkownik obsługiwany jest przez jedną ze stacji bazowych. Typowo jest to stacja od której odbiera najsilniejszy sygnał referencyjny. W tym przypadku przydział zasobów radiowych oraz całość przetwarzania sygnału (kodowanie kanałowe, prekodowanie, modulacja etc.) odbywa się niezależnie w obrębie każdej komórki. Nawet jeśli specyfikacje techniczne [4] umożliwiają pewną współpracę pomiędzy komórkami, np. w celu koordynacji interferencji, dotyczy ona tylko kilku stacji bazowych. Ponadto utworzony w ten sposób klaster stacji bazowych obsługujących użytkownika nie może być dynamicznie modyfikowany, przez co pokrycie sygnałem radiowym obszarów znajdujących się na granicy klastra jest zdecydowanie mniejsze niż w jego centrum. W związku z tym w systemach o architekturze zorientowanej na sieć występuje silna dysproporcja pomiędzy przepływnościami osiąganymi przez użytkowników znajdujących się na skraju komórki i tymi będącymi blisko jej centrum. Architektura zorientowana na użytkownika zakłada zgodnie z definicją podaną w [7], że użytkownik jest obsługiwany jednocześnie przez wszystkie stacje bazowe. Co więcej, sygnał nadawany przez wiele stacji bazowych jest odbierany przez użytkownika w sposób koherentny. Stacje bazowe podłączone są do tzw. Centralnej Jednostki Przetwarzania CPU (*ang. Central Processing Unit*). W zależności od implementacji podział przetwarzania sygnału pomiędzy stacje bazowe a CPU może się różnić. W przypadku pełnej centralizacji całość przetwarzania realizowana jest w CPU, a poszczególne stacje bazowe pełnią tylko funkcje tzw. zdalnych jednostek radiowych RRU (*ang. Remote Radio Unit*). Rozproszone implementacje przesuwają cześć przetwarzania sygnału do stacji bazowych np. prekodowanie sygnału, modulacja, kodowanie kanałowe. Zaletą implementacji scentralizowanej jest możliwość holistycznej optymalizacji transmisji w systemie, np. koordynacji interferencji. Wadą natomiast jest konieczność przesyłania dużej ilości danych kontrolnych i informacji, np. o współczynnikach kanałów radiowych.



Rysunek 1: Architektury systemu M-MIMO.

Dla pojedynczego bloku zasobów o indeksie n, moc sygnału odbieranego przez k-tego użytkownika w systemie M-MIMO zorientowanym na użytkownika, składającym się z N_{bs} stacji bazowych (często nazywanych również punktami dostępu (*ang. Access Points*) dana jest wzorem [9]:

$$P_{k,n} = \left| \sum_{l=1}^{N_{\text{bs}}} \sqrt{p_{kln}} \boldsymbol{h}_{kln}^{H} \boldsymbol{w}_{kln} \right|^2, \qquad (1)$$

gdzie p_{kln} jest mocą nadawaną w pasmie n-tego bloku zasobów przez stację bazową l do k-tego użytkownika, $\boldsymbol{h}_{kln}$ jest wektorem zespolonych współczynników kanału radiowego pomiędzy k-tym użytkownikiem, a każdą z M_l anten stacji bazowej l, stałych w pasmie n-tego bloku zasobów, a $\boldsymbol{w}_{kln}$ jest wektorem wag prekodera. Można zauważyć, że sygnał nadawany przez wszystkie stacje bazowe może się koherentnie dodawać podczas odbioru przez użytkownika, co ma szczególne znaczenie w przypadku, gdy znajduje się on w podobnej odległości od kilku stacji bazowych (na granicy geograficznej komórki, w sensie systemu zorientowanego na sieć). Należy jednak zauważyć, że podobną własność ma całkowita moc interferencji generowanej przez innych użytkowników, która dana jest wzorem [9]:

$$I_{k,n} = \sum_{i=1}^{K_n} \left(\left| \sum_{l=1}^{N_{\text{bs}}} \sqrt{p_{iln}} \boldsymbol{h}_{kln}^{H} \boldsymbol{w}_{iln} \right|^2 \right) - P_{k,n}, \qquad (2)$$

gdzie K_n oznacza liczbę użytkowników, którym przydzielono n-ty blok zasobów. Jak już wspomniano interferencje można znacznie ograniczyć w przypadku scentralizowanej implementacji systemu M-MIMO zorientowanego na użytkownika.

3. WYZWANIA IMPLEMENTACYJNE

System M-MIMO o architekturze zorientowanej na użytkownika pozwala uzyskać bardziej równomierny rozkład przepływności osiąganych przez użytkowników, kosztem szeregu wyzwań stojących przed praktyczną implementacją takiej architektury przyszłych sieci mobilnych. Do najważniejszych z nich należą [2]:

Wirtualne Klastry Obsługi W praktyce często może okazywać się, że obsługa każdego użytkownika przez wszystkie stacje bazowe stwarza konieczność wymiany bardzo dużej ilości danych sygnalizacyjnych i znacznie zwiększa złożoność obliczeniową przetwarzania sygnału, np. konieczne jest wymiana informacji dotyczących kanałów radiowych użytkowników. W wielu przypadkach część stacji bazowych będzie miała niewielki wpływ na wzrost przepływności danego użytkownika, np. z powodu znacznej odległości od niego. Z tej perspektywy istotne jest tworzenie tzw. wirtualnych klastrów obsługi (*ang. Serving Clusters*), czyli podzbiór wszystkich stacji bazowych, które obsługują danego użytkownika.

Opóźnienia i synchronizacja W związku z tym, że poszczególne stacje bazowe jednocześnie obsługujące użytkowników są od siebie oddalone o nawet kilka kilometrów istotnym problem mogą się okazać opóźnienia oraz synchronizacja. Problem dotyczy zarówno wymiany informacji sygnalizacyjnej pomiędzy stacjami bazowymi a CPU, jak i warstwy fizycznej, np. dobranie wag prekodera w niektórych przypadkach będzie musiało uwzględniać brak synchronizacji fazy pomiędzy poszczególnymi stacjami bazowymi.

Alokacja Zasobów W systemach zorientowanych na użytkownika potrzebne jest inne podejście do przydziału zasobów radiowych niż w przypadku systemów o architekturze zorientowanej na sieć. Dla systemu zorientowanego na użytkownika alokacja zasobów odbywa się w ramach

całej sieci (lub wirtualnego klastra obsługi), a nie pojedynczej komórki. Ważnym zagadnieniem dotyczącym alokacji zasobów jest dynamiczne dobieranie użytkowników w grupy, które mogą podlegać multipleksacji przestrzennej (*ang. User Pairing*).

Efektywność Energetyczna Sieć zorientowana na użytkownika musi brać pod uwagę zużycie energii potrzebne do zapewnienia odpowiedniej jakości usług. Zarówno zasoby sprzętowe w CPU, jak i te zainstalowane na stacjach bazowych muszą się w inteligentny sposób skalować i dostosowywać do aktualnego ruchu w sieci. Efektywność energetyczna może być zwiększana, np. poprzez tworzenie wirtualnych klastrów obsługi w taki sposób, żeby niektóre stacje bazowe mogły być wprowadzane w stan uśpienia. Szczególne znaczenie w tym kontekście ma wykorzystanie metod uczenia maszynowego i sztucznej inteligencji [5].

4. OPRACOWANY SYMULATOR

Na potrzeby porównania systemów M-MIMO o architekturze zorientowanej na użytkownika, z tymi o architekturze zorientowanej na sieć opracowany został zaawansowany symulator komputerowy. Jest to symulator systemowy (*ang. system-level simulator*), który implementuje łącze w dół (*ang. Downlink*) w systemie OFDM składającym się z 6 stacji bazowych podłączonych do CPU. System zajmuje pasmo o szerokości 25 MHz zaalokowane wokół częstotliwości środkowej równej 3.6 GHz i podzielone na $N_{\mathrm{rb}} = 69$ bloków zasobów. W systemie znajduje się jedna stacja bazowa typu makro (oznaczona indeksem $l = 0$) zainstalowana na wysokości 45 m, której całkowita moc nadawana jest równa $P_0^{\mathrm{tx}} = 46$ dBm i jest wyposażona w macierz składającą się z $M_0 = 128$ anten (8 rzędów $\times$ 16 kolumn).Pozostałe 5 stacji bazowych to stacje typu mikro zainstalowane na wysokości 6 m. Każda z nich nadaje z mocą $P_l^{\mathrm{tx}} = 30$ dBm i jest wyposażona w dwie macierze antenowe o rozmiarze $M_l = 16$ (8 rzędów $\times$ 2 kolumny). Model ruchu wykorzystywany w symulatorze to tzw. model pełnego bufora (*ang. Full Buffer*), w którym użytkownikom rozdzielane są wszystkie zasoby radiowe dostępne w systemie. Przydział zasobów radiowych odbywa się w formie scentralizowanej niezależnie dla każdego bloku zasobów, według procedury maksymalizującej tzw. metrykę PF (*ang. Proportional Fairness*) [8]. W ramach każdego bloku zasobów użyty algorytm sprawdza poziom korelacji kanałów radiowych użytkowników i na tej podstawie dokonuje multipleksacji w dziedzinie przestrzennej. Moc nadawana przez każdą stację bazową do zaalokowanych użytkowników zależy od ich warunków propagacyjnych i dana jest wzorem:

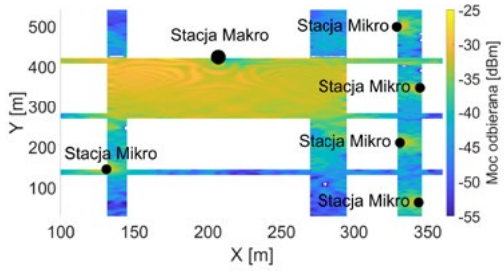
$$p_{kln} = \frac{P_l^{\mathrm{tx}}}{N_{\mathrm{rb}}} \cdot \frac{||\boldsymbol{h}_{kln}||^2}{\sum_{i=1}^{K_n} ||\boldsymbol{h}_{iln}||^2}, \tag{3}$$

gdzie $||\cdot||$ oznacza normę l^2 z wektora. W ten sposób żadna stacja bazowa nie będzie nadawała dużej mocy do użytkowników znajdujących się daleko od niej. Wszystkie stacje bazowe wykorzystują prekoder MRT (*ang. Maximum Ratio Transmission*), który dla każdego użytkownika maksymalizuje jego współczynnik sygnału do szumu:

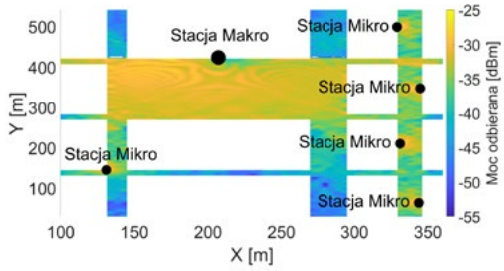
$$\boldsymbol{w}_{kln} = \frac{\boldsymbol{h}_{kln}}{||\boldsymbol{h}_{kln}||}. \tag{4}$$

W symulatorze zaimplementowanych zostało 15 schematów modulacji i kodowania MCS (*ang. Modulation and Coding Schemes*), dobieranych według procedury opisanej w [3], na podstawie stosunku sygnału do szumu i interferencji (obserwowanych i raportowanych przez użytkowników, oraz estymowanych na etapie przydziału zasobów radiowych) obliczonego z użyciem wzorów (1) i (2). Zaimplementowane środowisko symulacyjne wykorzystuje realistyczny model kanału radiowego bazujący na trójwymiarowym śledzeniu promieni dostarczony przez firmę Wireless InSite$^{\mathrm{TM}}$. Kanał radiowy został wygenerowany w scenariuszu miejskim zdefiniowanym w ramach projektu METIS, i jest to tzw. model *Madrid Grid* [6]. Każda ścieżka obliczona jest przy założeniu maksymalnie 15 odbić i jednej dyfrakcji.

5. WYNIKI SYMULACJI



(a) Systemu o architekturze zorientowanej na sieć



(b) System o architekturze zorientowanej na użytkownika

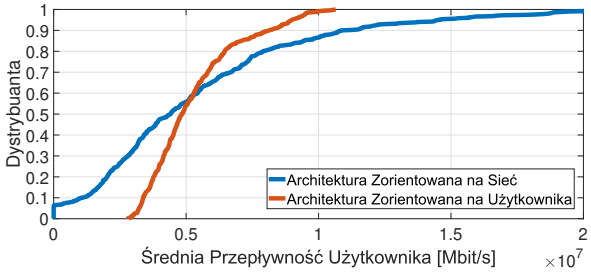
Rysunek 2: Maksymalna moc sygnału odbieranego

W ramach eksperymentu symulacyjnego, najpierw porównane zostało pokrycie sygnałem radiowym rozważnego obszaru w przypadku architektury zorientowanej na sieć i na użytkownika. W tym celu wygenerowany został kanał radiowy dla użytkowników rozmieszczonych równomiernie na obszarze sieci w odstępach co 5 m (łącznie 3542 użytkowników). Dla każdego użytkownika z użyciem wzoru (1) i przy założeniu prekodera MRT obliczony został poziom odbieranej mocy sygnału w przypadku gdy obsługuje go jedna stacja bazowa, oraz w przypadku sieci zorientowanej na użytkownika, kiedy obsługiwany jest przez wszystkie stacje bazowe. Wyniki uśrednione po wszystkich blokach zasobów zostały przedstawione odpowiednio na Rys. 2a i Rys. 2b, gdzie biały kolor oznacza budynki. Zalety zastosowania architektury zorientowanej na użytkownika są szczególnie widoczne w alejce z prawej strony, pokrytej przez liczne stacje mikro. Dla architektury zorientowanej na użytkownika alejka jest pokryta równomiernie sygnałem radiowym o mocy oscylującej wokół -35 dBm. W przypadku architektury zorientowanej na sieć widoczne są wyraźne granice pomiędzy komórkami, gdzie moc odbierana spada poniżej -50 dBm. Dzięki zastosowaniu architektury zorientowanej na użytkownika na granicach komórek użytkownicy mogą odbierać moc wyższą o na-

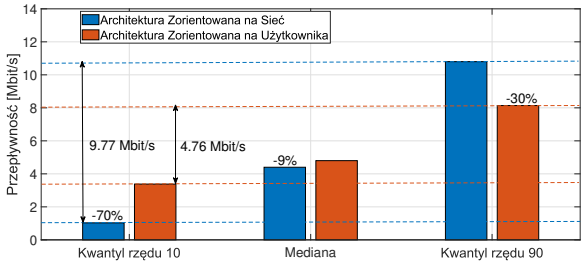
wet 15 dB niż w przypadku architektury zorientowanej na sieć. Następnie przeprowadzone zostały symulacje systemowe z użyciem środowiska opisanego w Rozdziale 4. Przeprowadzono 10 przebiegów symulacyjnych odpowiadających 0.5 s działania rzeczywistego systemu 5G (1000 slotów trwających 0.5 ms każdy). W ramach każdego przebiegu symulacyjnego rozważonych jest 50 użytkowników losowo rozmieszczonych na obszarze sieci i poruszających się w losowych kierunkach z prędkością 1.5 m/s. Na koniec każdego przebiegu symulacyjnego zebrane zostały statystyki dotyczące średniej przepływności każdego z rozważanych użytkowników. Dane o średniej przepływności użytkowników zebrane podczas 10 przebiegów symulacyjnych posłużyły do stworzenia rozkładu średnich przepływności użytkowników. Eksperyment symulacyjny w tych samych warunkach (dla tych samych jąder generatorów liczb losowych) przeprowadzono rozważając architekturę zorientowaną na użytkownika i na sieć. Na Rys. 3 porównane zostały dystrybuanty rozkładu średniej przepływności użytkowników w obu rozważanych architekturach systemu M-MIMO. Widać wyraźnie, że w przypadku architektury systemu zorientowanej na użytkownika przepływności osiągane przez użytkowników są bardziej równomierne niż w przypadku architektury zorientowanej na sieć. Ponadto widać dużą poprawę przepływności dla użytkowników mających najbardziej wymagające warunki radiowe. Na Rys. 4 porównane zostały kwantyle rzędu 10, 50 (mediana) i 90 z rozkładu średnich przepływności użytkowników. Architektura zorientowana na użytkownika pozwala poprawić medianę i kwantyl rzędu 10 odpowiednio o 9% i ponad trzykrotnie względem architektury zorientowanej na sieć. Warto zauważyć, że w przypadku architektury systemu M-MIMO zorientowanej na sieć występują duże dysproporcje pomiędzy kwantylem rzędu 10 a 90, tj. 9.77 Mbit/s, dzięki zastosowaniu architektury zorientowanej na użytkownika można zmniejszyć te dysproporcje o połowę do poziomu 4.76 Mbit/s. Kosztem wynikającym z zastosowania architektury zorientowanej na użytkownika jest pogorszenie przepływności osiąganych przez 10% użytkowników charakteryzujących się najlepszymi warunkami radiowymi (kwantyl rzędu 90) o 30%.

6. WNIOSKI

Nowoczesna architektura systemu M-MIMO pozwala jednocześnie obsługiwać użytkownika przez wszystkie stacje bazowe w systemie. Dzięki temu możliwe jest bardziej równomierne pokrycie sygnałem sieci, które przekłada się na znaczną poprawę przepływności osiąganych przez użytkowników charakteryzujących się najmniej korzystnymi warunkami radiowymi. Wyniki symulacyjne pokazały, że zysk może być ponad trzykrotny.



Rysunek 3: Porównanie dystrybuant rozkładu średnich przepływności użytkowników, dla systemu o architekturze zorientowanej na użytkownika i na sieć



Rysunek 4: Porównanie kwantyli 10,50 i 90 rozkładu średnich przepływności użytkowników, dla systemu o architekturze zorientowanej na użytkownika i na sieć

LITERATURA

[1] Akyildiz, Ian F., Kak, Ahan, Nie, Shuai, 2020, "6G and Beyond: The Future of Wireless Communications Systems". *IEEE Access*, 8: 133995–134030.

[2] Ammar, Hussein A., Adve, Raviraj, Shahbazpanahi, Shahram, Boudreau, Gary, Srinivas, Kothapalli Venkata, 2022, "User-Centric Cell-Free Massive MIMO Networks: A Survey of Opportunities, Challenges and Solutions". *IEEE Communications Surveys & Tutorials*, 24 (1): 611–652.

[3] Bossy, Bartosz, Kryszkiewicz, Paweł, Bogucka, Hanna, 2017, "Optimization of energy efficiency in the downlink lte transmission". *2017 IEEE International Conference on Communications (ICC)*, 1–6.

[4] Dahlman, Erik, Parkvall, Stefan, Sköld, Johan, 2018, "Chapter 4 - LTE—An Overview". Erik Dahlman, Stefan Parkvall, Johan Sköld (Editors), *5G NR: the Next Generation Wireless Access Technology*, 39–55, Academic Press, ISBN 978-0-12-814323-0.

[5] Ghiasi, Niyousha, Mashhadi, Shima, Farahmand, Shahrokh, Razavizadeh, S. Mohammad, Lee, Inkyu, 2023, "Energy Efficient AP Selection for Cell-Free Massive MIMO Systems: Deep Reinforcement Learning Approach". *IEEE Transactions on Green Communications and Networking*, 7 (1): 29–41.

[6] METIS, 2013, *Deliverable D6.1,Simulation guidelines v1.0.* Mobile and wireless communications Enablers for the Twenty-twenty Information .

[7] Ngo, Hien Quoc, Ashikhmin, Alexei, Yang, Hong, Larsson, Erik G., Marzetta, Thomas L., 2015, "Cell-free massive mimo: Uniformly great service for everyone". *2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 201–205.

[8] Yoo, Taesang, Goldsmith, A., 2006, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming". *IEEE Journal on Selected Areas in Communications*, 24 (3): 528–541.

[9] Özlem Tugfe Demir, Björnson, Emil, Sanguinetti, Luca, 2021, "Foundations of user-centric cell-free massive mimo". *Foundations and Trends® in Signal Processing*, 14 (3-4): 162–472.

NIEZAKŁÓCONA PRYWATNA SIEĆ 5G NA POTRZEBY SYTUACJI KRYZYSOWEJ

UNDISRUPTED PRIVATE 5G NETWORK IN CASE OF CRISIS SITUATION

Mateusz Kustra[1]; Bogusław Grochowina[2]; Krzysztof Kosmowski[3]

[1] Wojskowy Instytut Łączności - PIB, Zegrze Południowe, m.kustra@wil.waw.pl
[2] Wojskowy Instytut Łączności - PIB, Zegrze Południowe, b.grochowina@wil.waw.pl
[3] Wojskowy Instytut Łączności - PIB, Zegrze Południowe, k.kosmowski@wil.waw.pl

DOI: 10.15199/59.2023.4.14

Streszczenie: **W referacie przedstawiono propozycję wykorzystania prywatnych sieci 5G do realizacji szybkiej i niezawodnej łączności na danym terenie w sytuacji zarządzania kryzysowego. Założono wykorzystanie tych samych pasm częstotliwości co systemy komercyjne pracujące na danym terenie. Scharakteryzowano model systemu oraz omówiono rozpatrywane miary i kryteria zakłócalności. Przedstawiono również wyniki wykonanych za pomocą oprogramowania HTZ Warfare badań symulacyjnych dla zakładanego przypadku.**

Abstract: **The paper presents a proposal for utilizing private 5G networks to establish fast and reliable communication in a given area during crisis management situations. The assumption is to use the same frequency bands as the commercial systems operating in that area. The model of such a system was described, and the considered measures and criteria of interference were discussed. Additionally, the results of simulation studies conducted using HTZ Warfare software for the assumed scenario were presented.**

Słowa kluczowe: **5G, interferencje, miary i kryteria zakłócalności, sieć prywatna, współczynnik ochronny.**

Keywords: **5G, interferences, metrics and criteria of interference, private network, protection ratio.**

1. WSTĘP

Sieć 5G jest projektowana z myślą o podłączeniu do niej urządzeń tzw. Internetu rzeczy, które dzięki stałemu połączeniu i zastosowaniu algorytmów sztucznej inteligencji, umożliwią rozwinięcie inteligentnego przemysłu, inteligentnych miast czy inteligentnego transportu. Osiągnięcie takich funkcjonalności wymaga znaczącego przyrostu wartości parametrów względem wartości uzyskanych w systemach 4G [1].

Dzięki powyższym korzyściom, technologia 5G została uznana jako strategiczna w obszarach biznesowych oraz działalności przemysłowej (produkcyjnej, logistycznej, motoryzacyjnej). Sieci 5G mają umożliwić sprawne wejście w erę tzw. Przemysłu 4.0. Stąd niezwykle istotnym stają się prywatne sieci 5G, które będą wykorzystywane na terenach silnie uprzemysłowionych.

Jednakże szybka i niezawodna łączność jest również warunkiem efektywnego zarządzania w szeroko rozumianej sytuacji kryzysowej. W takim scenariuszu prywatna sieć 5G również będzie stanowiła kluczowy element.

O ile prywatne sieci 5G na potrzeby Przemysłu 4.0 budowane są w oparciu o wydzielone zasoby częstotliwościowe, o tyle w przypadku prywatnej sieci 5G postawionej ad-hoc na danym terenie w sytuacji kryzysowej wydzielonych częstotliwości może nie być.

Biorąc pod uwagę powyższe rozważania, ideą tego artykułu jest przedstawienie możliwości wdrożenia prywatnej sieci 5G na wybranym terenie w środowisku z już istniejącą infrastrukturą zrealizowaną w omawianej technologii. Założeniem jest również praca prywatnej sieci 5G w tym samym paśmie co sieć komercyjna.

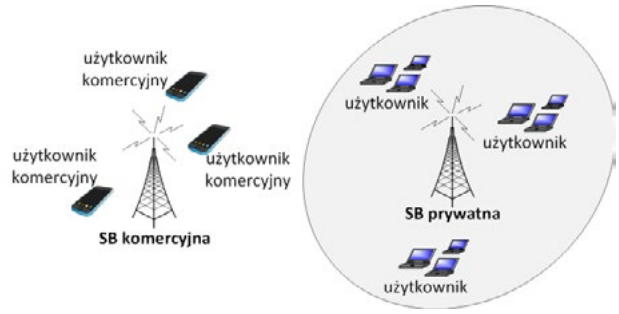
W związku z powyższym koniecznym jest sprawdzenie wpływu sieci komercyjnej na możliwość zapewnienia niezakłóconej transmisji na terenie przewidzianym dla sieci prywatnej. Brak analizy wpływu w drugą stronę, to jest sygnału sieci prywatnej na sygnał sieci komercyjnej podyktowany jest tym, że zakłada się pracę na terenie gdzie użytkowników komercyjnych jest relatywnie mało oraz priorytetowość łączności dla sieci prywatnej w sytuacji kryzysowej.

W dalszej części artykułu został omówiony rozpatrywany model systemu oraz zdefiniowane miary i kryteria zakłócalności. Następnie przedstawiono wyniki badań symulacyjnych wykonanych w oprogramowania HTZ Warfare. Uzyskane rezultaty pozwalają określić w jakiej odległości od komercyjnej stacji bazowej (SB) należy ulokować stację bazową prywatnej sieci 5G tak aby na danym obszarze móc odbierać sygnał bez interferencji.

2. MODEL SYSTEMU

W artykule rozpatrywana jest prywatna sieć 5G, której obszar działania ma znajdować się w promieniu do 1,5 km od stacji bazowej, której rozlokowanie zakłada się w niedalekiej odległości od stacji bazowej już istniejącego komercyjnego systemu 5G (Rys. 1).

W związku z powyższym należy założyć, że na sygnał sieci prywatnej na zakładanym obszarze będzie wpływał sygnał sieci komercyjnej. Stąd w badaniach będą brane różne odległości pomiędzy SB sieci prywatnej a SB sieci komercyjnej. Pierwsza rozpatrywana odległość to 2 km (Rys. 2).



Rys. 1. Model zakładanego systemu

Rys. 2. Zakładane rozmieszczenie SB dla sieci prywatnej i komercyjnej

W rozważanym modelu zakłada się, że system 5G dla sieci prywatnej jak i komercyjnej ma takie same parametry. Zdefiniowane w modelu podstawowe parametry zostały przedstawione w tabeli poniżej.

Tab. 1. Podstawowe parametry systemu [3]

Parametr	*Wartość*	*Jednostka*
Częstotliwość	3510	[MHz]
Pasmo kanału	20	[MHz]
Moc	6	[W]
Wysokość masztu antenowego	4	[m]
Wysokość anteny odbiorczej	1,5	[m]
Numerologia	1	-
Odstęp podnośnych	30	[kHz]
Modulacja	256 QAM	-

W powyższej tabeli znajduje się parametr o nazwie „numerologia" (ang. *numerologies*), który wymaga krótkiego wyjaśnienia. Numerologia jest to pojęcie związane z organizacją warstwy fizycznej 5G NR [4]. W pierwszym przybliżeniu należy stwierdzić, że jest to odstęp między podnośnymi OFDM, który może wynosić od 15kHz – do 960 kHz (w omawianym przypadku wynosi on 30kHz). Jednak ma on głębsze znaczenie, ponieważ wpływa również na organizację kanału radiowego w dziedzinie czasu. W zależności od odstępu między podnośnymi zmienia się czas trwania symbolu OFDM, podczas gdy struktura czasowa zorganizowana w formę ramek (ang. *frames*), podramek (ang. *subframes*) jest stała. Stąd też liczba symboli OFDM w ramce jest zmienna i zależy od przyjętej numerologii.

3. MIARY I KRYTERIA ZAKŁÓCALNOŚCI

Podstawowym kryterium kompatybilności międzysystemowej (zgodnie z [5]) są tzw. współczynniki ochronne (ang. *protection ratio* – PR). Definiują one taki stosunek mocy sygnału użytecznego do zakłócającego (w dB), przy którym występujące zakłócenia nie powodują niedopuszczalnej degradacji jakości danych usług. Zamiennie oznacza się go również jako minimalny wymagany stosunek C/I (*carrier-to-interference*).

W przypadku służb radiodyfuzyjnych stałych takich jak np. DVB-T współczynniki ochronne zwykle wyznacza się na podstawie pomiarów laboratoryjnych przeprowadzanych przez wiele niezależnych instytucji. Następnie wyniki badań uzgadnia się wspólnie na forum międzynarodowym w ramach prac CEPT (ang. *European Conference of Postal and Telecommunications Administrations* – Europejska Konferencja Administracji Poczty i Telekomunikacji) i ITU (ang. *International Telecommunication Union* – Międzynarodowy Związek Telekomunikacyjny).

Możliwe jest także analityczne wyznaczanie współczynników ochronnych. W tym celu niezbędne jest opracowanie analitycznego modelu odbiornika danego systemu i metody analizy kompatybilności. Dodatkowo przy wyznaczaniu PR należy wziąć pod uwagę zjawiska zaprezentowane w literaturze przedmiotu [6].

Zgodnie z [6] współczynnik ochronny przyjmuje następującą postać:

$$PR = {}^{S}/_{N_{(BER\ 10^{-Y})}} + {}^{N}/_{I} + MIA + FM - NFD \quad (1)$$

gdzie:

- S/N – stosunek sygnał/szum, dla którego osiągana jest wymagana bitowa stopa błędów, w przypadku systemów cyfrowych oznaczany również jako C/N (ang. *carrier-to-noise*). Poziom ten jest związany z czułością systemu odbiorczego;
- N/I – stosunek szum/interferencje (taki jak 10, 6 lub 3 dB), który powoduje ok. 0.5, 1 lub 2 dB degradację poziomu sygnału użytecznego z powodu interferencji;
- MIA –możliwość wystąpienia wielokrotnych interferencji;
- FM – margines na zaniki (ang. *fade margin*) jest to czynnik, który bierze pod uwagę fakt, że poziom sygnału użytecznego fluktuuje ze względu na zmienne warunki propagacji. Jego wartość może się znacząco różnić w zależności od rodzaju rozpatrywanego systemu. W przypadku gdy rozpatrywany jest model kanału AWGN ten składnik może zostać pominięty;
- NFD – dyskryminacja częstotliwościowa tzw. filtra sieciowego wynikającego z maski nadajnika i selektywności odbiornika.

W celu wykonania badań, wartości współczynników ochronnych zostały wyznaczone przy założeniu następujących wartości parametrów występujących we wzorze (1):

- N/I = 3 dB;
- FM = 3 dB;
- MIA = 6 dB;

oraz dla Noise Figure = 8 dB.

Parametry te zostały przyjęte arbitralnie i ich zmiana będzie skutkować zmianą wartości wyznaczonych współczynników PR. Natomiast współczynnik S/N został wyznaczony na podstawie tabel z dokumentu ETSI [2].

Obliczenia zostały wykonane dla wartości SCS (ang. *Subcarrier Spacing*) 15, 30, 60 kHz. Ze względu na mnogość otrzymanych wyników, poniżej zamieszczony został tylko wycinek jednej z tabel z wynikami obliczeń

Tab. 2. Współczynniki ochronne dla UE, odstęp między podnośnymi OFDM 15kHz

Pasmo pracy	*Szerokość kanału [MHz]*				
	5	*10*	*15*	*20*	···
n1	11,47	11,49	11,47	11,39	···
n2	13,47	13,49	13,47	13,39	···
n3	14,47	14,49	14,47	14,39	···
n5	13,47	13,49	13,47	14,39	···
n7	13,47	13,49	13,47	13,39	···
n8	14,47	14,49	14,47	15,19	···
...	...	...	...	...	...

Przyjmując, że dla rozpatrywanego systemu zostały określone wartości współczynnika ochronnego zdefiniowanego jako minimalny, względny stosunek mocy sygnału użytecznego i niepożądanego na wejściu zakłócanego odbiornika – jego wartość należy porównać z wartością C/I. Stosunek mocy C/I powinien być większy niż wartość współczynnika ochronnego PR zgodnie z (2).

$$C/I \geq PR \tag{2}$$

4. BADANIA SYMULACYJNE

Jakość sygnału docierającego od nadajnika do odbiornika jest uzależniona od charakterystyk propagacyjnych środowiska, w którym rozchodzą się fale radiowe. Charakterystyki te podlegają znaczącym zmianom w zależności od topologii terenu (płaski, otwarty, zabudowany, zalesiony itp.), wykorzystywanych częstotliwości, mobilności terminali, czy też poziomu interferencji.

W badaniach jakości sygnału sieci prywatnej będą brane pod uwagę interferencje wprowadzane przez system komercyjny.

Badania symulacyjne zostały przeprowadzone w narzędziu symulacyjnym HTZ Warfare. Oprogramowanie to dostarcza wszystkich niezbędnych funkcji do projektowania, optymalizacji i oceny radiowych lub mikrofalowych łączy naziemnych, morskich, lotniczych i lotniczo-naziemnych.

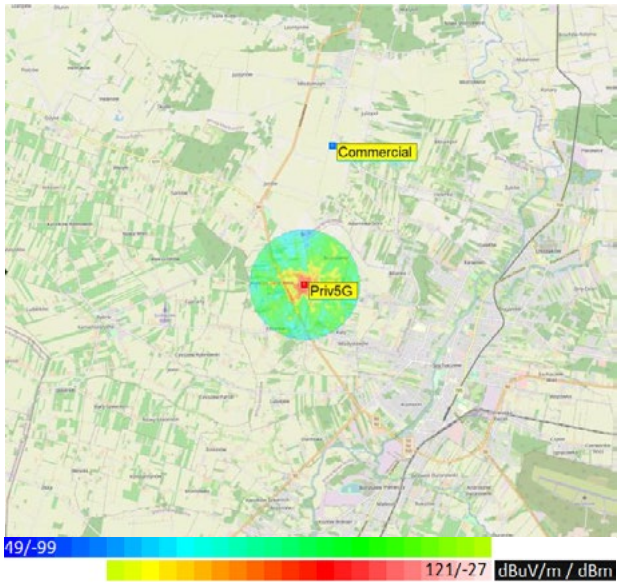
HTZ Warfare umożliwia również analizę interferencji w badanym systemie łączności. Jedną z metod analizy jest analiza *Interference on best server*, która umożliwia generację mapy interferencji obszaru nadajnika posiadającego kategorię *best server* (nadajnika dominującego na danym obszarze). Razem z mapą generowana jest tablica, określająca jaka część obszaru pokrycia danego nadajnika kategorii *best server* jest zakłócona przez inne nadajniki występujące na danym obszarze.

W celu oceny niezakłóconej pracy systemu 5G przyjęto powyższe dane jako miarę interferencji. Przyjęto, że system jest wolny od interferencji jeżeli część obszaru pokrycia nadajnika kategorii *best server* zakłócona przez inne nadajniki jest mniejsza od 1%.

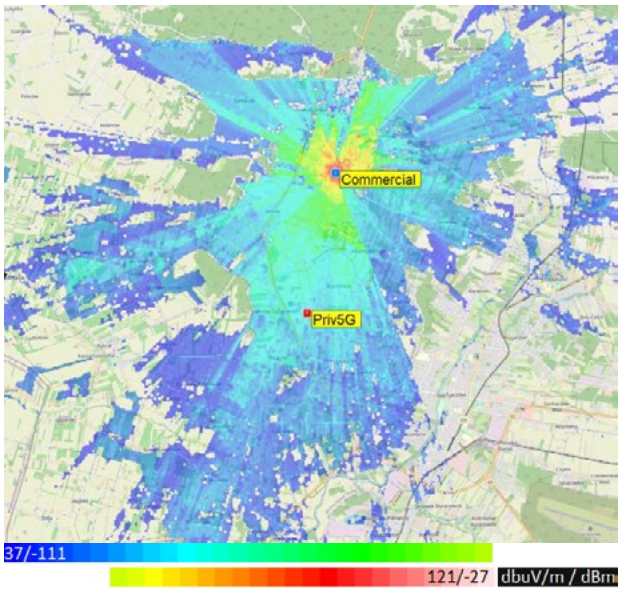
Dodatkowo w celu analizy interferencji (jako kryterium/warunek wykrycia interferencji) przyjęto wyliczoną zgodnie rozdziałem 3 wartość współczynnika protekcji wynoszącą 11,39 dB oraz odległość między stacjami bazowymi.

4.1. Analiza pokrycia obszaru sieci sygnałem stacji bazowych

W procesie projektowania systemu łączności radiowej istotnym czynnikiem jest to czy dany teren zostanie pokryty sygnałem stacji bazowej. W związku z tym w pierwszej kolejności przeprowadzono badania symulacyjne mające na celu określenie współczynnika pokrycia terenu (ang. *coverage factor*) dla SB sieci prywatnej w zakładanej odległości 1,5 km (Rys. 3). Jak widać na całym obszarze powinna być możliwa łączność w ramach sieci prywatnej.



Rys. 3. Mapa zasięgu SB sieci prywatnej w zakładanym obszarze



Rys. 4. Mapa zasięgu SB komercyjnej

Kolejnym krokiem było sprawdzenie zasięgu łączności dla SB komercyjnej (Rys. 4).

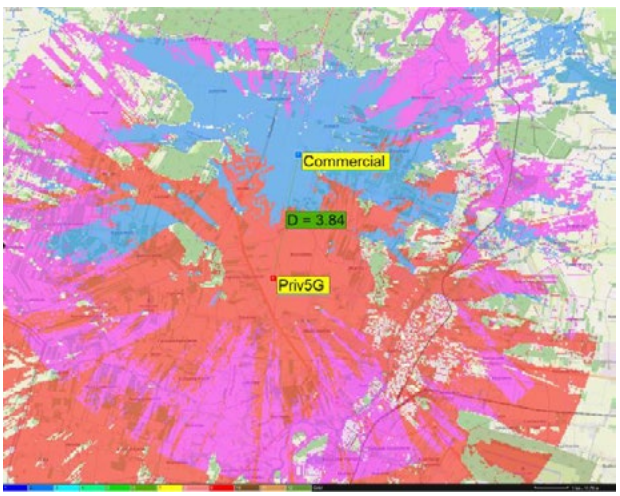
Badanie wykazało, że cały obszar sieci prywatnej znajduje się w zasięgu SB komercyjnej. Zachodzi pytanie czy sygnał SB komercyjnej będzie miał wpływ na łączność w ramach rozpatrywanej sieci prywatnej. W tym celu została przeprowadzona analiza interferencji (4.2).

4.2. Analiza interferencji

Analiza interferencji miała wykazać w jakiej minimalnej odległości musi zostać ustawiona SB sieci prywatnej od SB komercyjnej, aby na zakładanym obszarze sygnał SB komercyjnej nie wprowadzał szkodliwych interferencji. W tym celu zostały wygenerowane mapy interferencji dla różnych odległości pomiędzy SB sieci prywatnej a komercyjnej.

Na Rys. 5 i Rys. 6 przedstawiono mapy interferencji dla dwóch położeń SB sieci prywatnej względem SB komercyjnej. Kolor czerwony przedstawia obszary pokrycia terenu niezakłóconym sygnałem SB sieci prywatnej, kolor niebieski dla SB komercyjnej, natomiast obszary w których będą występowały niepożądane interferencje mają kolor różowy.

Rys. 5. Mapa interferencji w odległości 2,26 km



Rys. 6. Mapa interferencji w odległości 3,84 km

Przeprowadzona analiza interferencji wykazała, że umiejscowienie SB sieci prywatnej w odległości minimum 3,8 km od SB komercyjnej pozwoli zapewnić niezakłóconą łączność na żądanym obszarze dla sieci prywatnej. W zakładanym scenariuszu, każda mniejsza odległość spowoduje zakłócenia w odbiorze sygnału sieci prywatnej.

5. PODSUMOWANIE

W referacie zaproponowano wykorzystanie prywatnych sieci 5G do realizacji szybkiej i niezawodnej łączności na danym terenie w sytuacji zarządzania kryzysowego. Jest to istotny aspekt z punktu widzenie zarządzania widmem częstotliwości, ponieważ w przypadku sytuacji kryzysowej, gdy liczy się czas reakcji może nie być możliwości uzyskania pozwolenia na wykorzystanie wolnego pasma częstotliwości na danym terenie.

W związku z powyższym w artykule zaproponowano model systemu, który odpowiadałby ww. scenariuszowi oraz scharakteryzowano rozpatrywane miary i kryteria zakłócalności. Następnie przedstawiono wyniki badań symulacyjnych wykonanych w oprogramowania HTZ Warfare. Przeprowadzona analiza wykazała, że SB prywatnej sieci 5G powinna być oddalona na odległość ok. 3,8 km od SB komercyjnej, aby na danym obszarze móc odbierać sygnał bez interferencji. Ma to umożliwić realizację niezakłóconej pracy środków łączności systemu 5G na potrzeby efektywnego zarządzania kryzysowego.

6. LITEREATURA

[1] Dahlman Erik, Parkvall Stefan, Skold Johan, 2018. „5G NR The next generation wireless access technology", *Elsevier Ltd.*

[2] ETSI TS 138 101-1 V15.2.0 (2018-07). 5G; NR; User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone (3GPP TS 38.101-1 version 15.2.0 Release 15)

[3] ETSI TS 138 104 V15.3.0 (2018-10). 5G; NR; Base Station (BS) radio transmission and reception (3GPP TS 38.104 version 15.3.0 Release 15)

[4] ETSI TS 138 211 V17.1.0 (2022-04). 5G; NR; Physical channels and modulation *(*3GPP TS 38.211 version 17.1.0 Release 17*)*

[5] „Metody badania wpływu zakłóceń systemów radiowych następnych generacji (LTE, IEEE 802.22, DAB+, DVB-T) na istniejące środowisko radiowe w celu efektywnego gospodarowania widmem", 2010, Raport Z21/21300020/1392/10, Instytut Łączności Państwowy Instytut Badawczy

[6] Suh Kyoung Whoan, 2007. „A generalized formulation of the protection ratio applicable to frequency coordination in digital radio relay networks" *Radio Science*, 42: 1-10.

WYKORZYSTANIE REKONFIGUROWALNYCH INTELIGENTNYCH MATRYC ANTENOWYCH W ŁĄCZU DOSYŁOWYM SIECI 5G/6G WYKORZYSTUJĄCEJ BEZZAŁOGOWE STATKI POWIETRZNE

ON THE USE OF RECONFIGURABLE INTELLIGENT SURFACES FOR BACKHAULING IN UAV-ASSISTED 5G/6G NETWORK

Salim Janji[1]; Paweł Sroka[2]; Adrian Kliks[3]

[1] Politechnika Poznańska, Intytut Radiokomunikacji, Poznań, salim.janji@doctorate.put.poznan.pl
[2] Politechnika Poznańska, Intytut Radiokomunikacji, Poznań, pawel.sroka@put.poznan.pl
[3] Politechnika Poznańska, Intytut Radiokomunikacji, Poznań, adrian.kliks@put.poznan.pl

DOI: 10.15199/59.2023.4.15

Streszczenie: **Drony, dzięki możliwości ich szybkiego rozmieszczenia w trudnym terenie, uważane są za jeden z kluczowych elementów systemów bezprzewodowych 6G. Jednak w celu wykorzystania ich jako punkty dostępowe sieci konieczne jest zapewnienie łącza dosyłowego o odpowiedniej przepustowości. Dlatego w niniejszym artykule rozważane jest zwiększenie zasięgu sieci bezprzewodowej przez zapewnienie łącza dosyłowego dla końcowego punktu dostępowego z wykorzystaniem określonej liczby dronów-przekaźników oraz rekonfigurowalnych inteligentnych matryc antenowych (RIS). Zaprezentowane wyniki badań symulacyjnych pokazują, że użycie RIS pozwala na znaczące zwiększenie zasięgu sieci bez konieczności stosowania dodatkowych przekaźników.**

Abstract: **Unmanned Aerial Vehicles, due to the possibility of their fast deployment, are considered an essential element of the future wireless 6G communication systems. However, an essential enabler for their use as access points is to provide a sufficient throughput wireless backhaul link. Thus, in this paper we consider the aspect of extension of network coverage with the use of drone-based relaying and reconfigurable intelligent surfaces (RIS) for backhauling. Presented results of simulation experiments indicate that the use of RIS allows for significant improvement of network coverage without the need to use additional relays.**

Słowa kluczowe: **łącze dosyłowe, przekaźniki, drony, RIS, 6G**

Keywords: **backhaul, relays, RIS, UAV, 6G**

1. WSTĘP

Jednym z podstawowych zastosowań systemów bezprzewodowych piątej generacji (5G) jest zapewnienie usług szerokopasmowej transmisji danych (ang. *enhanced mobile broadband* – eMBB). Rozwiązaniem pozwalającym na realizację tego zadania jest rozmieszczenie bardzo dużej liczby stacji bazowych (ang. *base station* – BS) i punktów dostępowych (ang. *access point* – AP), tworząc w ten sposób sieć o bardzo dużym zagęszczeniu (ang. *ultra-dense network* – UDN). Jednak użycie tak dużej liczby BS i AP wymaga zapewnienia odpowiedniej jakości łączy dosyłowych w celu wymiany danych z siecią rdzeniową [1], co może okazać się niemożliwe lub nieopłacalne w przypadku stosowania łączy przewodowych.

Rozwiązaniem problemu zapewnienia łącza dosyłowego w UDN może być użycie komunikacji bezprzewodowej oraz bezzałogowych statków powietrznych (ang. *unmanned aerial vehicle* –UAV), czyli dronów, działających jako przekaźniki radiowe [2]. Rozmieszczenie UAV w sposób pół-statyczny pozwala utworzyć sieć tymczasowych łączy dosyłowych między stacją bazową makro i AP tworzącymi małe komórki. Dzięki odpowiedniemu zlokalizowaniu dronów-przekaźników (ang. *drone relay station* – DRS) możliwe jest uzyskanie łączy z bezpośrednią widocznością (ang. *line-of-sight* –LoS), co zapewnia większą przepustowość niż w przypadku łączy bez bezpośredniej widoczności.

Zagadnienie rozmieszczenia DRS i sterowania bezprzewodowym łączem dosyłowym z wieloma przeskokami jest skomplikowane, gdyż wpływa na nie szereg ograniczeń, takich jak: rozmiar UAV, jego mobilność, zużycie energii czy wolumen przesyłanych danych. Dodatkowo, w środowisku miejskim nie zawsze możliwe jest zapewnienie łączności typu LoS z ograniczoną liczba dronów. Dlatego jednym z rozwiązań wspomagających użycie UAV może być zastosowanie rekonfigurowalnych inteligentnych matryc antenowych (ang. *reconfigurable inteligent surfaces* – RIS), pozwalających na stworzenie wirtualnych połączeń LoS między nadajnikiem i odbiornikiem, dla których w normalnych warunkach bezpośrednia widoczność nie występuje [3]. RIS to matryca zawierająca dużą liczbę elementów, których zadaniem – w najprostszym ujęciu - jest odbicie docierającego sygnału radiowego, jednocześnie zmieniając jego amplitudę, częstotliwość lub fazę [4]. Dzięki użyciu RIS pojawia się możliwość oddziaływania na propagację sygnału zwiększając w ten sposób przepustowość, niezawodność czy zasięg transmisji. Zaletą RIS jest łatwość ich rozmieszczenia poprzez montaż na ścianach budynków, bilbordach, pojazdach czy statkach powietrznych. Dodatkowo są one efektywniejsze energetycznie od tradycyjnych przekaźników i w pełni kompatybilne z istniejącymi systemami komunikacji bezprzewodowej. W przypadku wykorzystania RIS w komunikacji z UAV, pozwalają one na utworzenie krótszych, niezawodnych wirtualnych łączy LoS, w szczególności w środowisku miejskim. Łączna optymalizacja położenia UAV i alokacja zasobów radiowych uwzględniająca użycie RIS pozwala też na minimalizację zużycia energii w dronach. Zagadnienie wykorzystania RIS w realizacji łącza dosyłowego z wykorzystaniem dronów jest stosunkowo nowym rozwiązaniem, w związku z

czym dostępne są jedynie nieliczne publikacje. Rozwiązanie zaprezentowane w [5] wykorzystuje matryce umieszczone na statkach powietrznych dużych wysokości, optymalizując łącze dosyłowe w zakresie efektywności energetycznej przez podział matrycy i odpowiednie fazowanie jej elementów. Z kolei w [6] rozważono RIS zamontowane na dronach, stosując model wielorękiego bandyty do optymalizacji łącza dosyłowego w paśmie fal milimetrowych.

W niniejszej pracy, rozwijając badania przedstawione w [7], rozważany jest problem rozmieszczenia DRS i wyboru ścieżki transmisji od stacji bazowej makro do AP w celu realizacji łącza dosyłowego w środowisku miejskim, gdzie drony znajdują się poniżej wysokości budynków. Realizacja łącza odbywa się z wieloma przeskokami z wykorzystaniem pasma fal milimetrowych i określonej minimalnej przepustowości. Zakładane jest też wykorzystanie RIS rozmieszczonych na ścianach wybranych budynków, co pozwala na zwiększenie zasięgu sieci przy użyciu stałej liczby DRS.

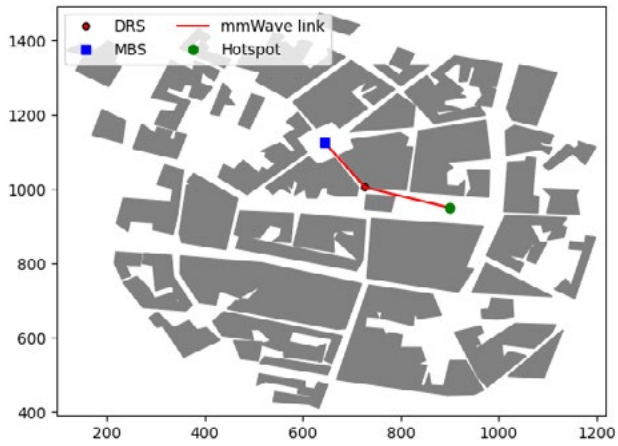
Dalsza część artykułu przedstawia się następująco: w rozdziale 2. przedstawiono model rozważanego systemu oraz sformułowano problem optymalizacyjny. W rozdziale 3 przedstawiono proponowane rozwiązanie oraz wyniki badań symulacyjnych jego zastosowania. Rozdział 4. podsumowuje wyniki dotychczasowych badań.

2. MODEL SYSTEMU I PROBLEM OPTYMALIZACYJNY

2.1. Model systemu

W niniejszej pracy rozważane jest zagadnienie wykorzystania UAV dla realizacji bezprzewodowego łącza dosyłowego z wieloma przeskokami w środowisku miejskim, gdzie dostępna jest tylko jedna stacja bazowa makro (MBS) odpowiedzialna za łączność z siecią rdzeniową. W związku z tym poszerzenie zasięgu sieci realizowane jest z wykorzystaniem mobilnych punktów dostępowych (AP) w postaci dronów połączonych bezprzewodowo z MBS. Zakładamy dostępność pasma o szerokości B w zakresie częstotliwości fal milimetrowych w celu realizacji łącza dosyłowego, a także wykorzystanie maksymalnie N dronów, które mogą działać jako AP lub DRS. Rozważany obszar miejski został zaprojektowany na podstawie analizy fragmentu zabudowy miasta Poznań pobranego z serwisu OpenStreetMap, co przedstawiono na rys. 1. Drony, po określeniu ich oczekiwanej lokalizacji, działają w sposób pół statyczny – nie zmieniają swojego położenia w trakcie realizacji transmisji. Dodatkowo, w systemie możliwe jest zastosowanie do R matryc RIS umieszczonych na wybranych ścianach budynków w celu poprawy zasięgu pojedynczego przeskoku łącza dosyłowego.

Ze względu na wymagania minimalnej przepustowości tworzonego łącza (C_{min}) zakłada się, że łącze dosyłowe dla wybranej lokalizacji AP jest dostępne jeśli dla każdego z przeskoków, tj. dla każdego łącza między MBS i DRS, dwoma DRS i między DRS i AP, można osiągnąć założoną minimalną przepustowość. Dodatkowo ze względu na wykorzystanie częstotliwości fal milimetrowych założono, że straty wynikające z odbić fal od obiektów innych niż RIS są na tyle wysokie, że nie pozwalają na realizację łącza dosyłowego, w związku z czym wymagane jest zawsze łącze LoS. Natomiast możliwa jest



Rys. 1. Ilustracja rozważanego scenariusza - schemat układu budynków, stacji bazowej makro (MBS) i rozważanej lokalizacji DRS i AP (hotspot) – opracowany na podstawie struktury zabudowy miasta Poznań

transmisja w konfiguracji wirtualnego LoS, tzn. wykorzystując odbicie sygnału od RIS - w tym przypadku całkowite tłumienie takiego łącza jest sumarycznym tłumieniem dwóch ścieżek: od nadajnika do RIS i od RIS do odbiornika. W związku z tym tłumienie pojedynczego przeskoku łącza może być wyznaczone następująco:

- w przypadku łącza bezpośredniego LoS (bez RIS) [8]:

$$\mathrm{PL}_{\mathrm{direct}}{}^{\mathrm{dB}}(d) = \mathrm{PL}(d_0) + 10\alpha \cdot \log(d) \qquad (1)$$

gdzie d to odległość między nadajnikiem i odbiornikiem, $\mathrm{PL}(d_0)$ to tłumienie wolnej przestrzeni dla $d_0 = 5\ m$, a α to współczynnik tłumienia ścieżki (wykładnik). W niniejszej pracy założono $\mathrm{PL}(d_0) = 39$ dB dla transmisji w paśmie 38 GHz i $\alpha = 2.13$ [9].

- dla przeskoku z wykorzystaniem RIS [10]:

$$\mathrm{PL}_{\mathrm{RIS}}{}^{\mathrm{dB}}(d_{1\to R}, d_{R\to 2}) = \mathrm{PL}(d_0) + 10\beta \cdot \log(M^2(d_{1\to R} + d_{R\to 2})) - g_{bf} \qquad (2)$$

gdzie M to liczba meta-powierzchni w RIS, β to współczynnik tłumienia ścieżki (wykładnik), a $d_{1\to R}$ i $d_{R\to 2}$ to odpowiednio odległości między nadajnikiem i RIS oraz między RIS i odbiornikiem. Dodatkowo założono, że RIS realizuje odbicie sygnału w sposób aktywny, zapewniając jego wzmocnienie g_{bf}, dzięki zastosowaniu formowania wiązki. W pracy założono $M = 3$ i $\beta = \alpha = 2.13$.

Średnią przepustowość wybranego łącza można wyznaczyć ze zmodyfikowanego wzoru Shannona [11]:

$$C_i = \eta \cdot B^{(\mathrm{eff})} \log_2(1 + \mathrm{SNR}_i), \qquad (3)$$

gdzie η oznacza efektywność łącza (udział bitów danych w całkowitej ich liczbie), $B^{(\mathrm{eff})}$ to efektywnie wykorzystane pasmo, a SNR_i to średni stosunek mocy sygnału do mocy szumu (SNR) i-tego przeskoku, wyznaczany jako:

$$\mathrm{SNR}_i = \frac{P_i^{(\mathrm{TX})} g_i}{\sigma^2}, \qquad (4)$$

oznaczając jako $P_i^{(\mathrm{TX})}$ i g_i odpowiednio moc nadawczą (ograniczoną jako $P_i^{(\mathrm{TX})} \le P_{\max}$) i wzmocnienie kanału dla i-tego przeskoku, a moc szumu jako σ^2. Wzmocnienie kanału wyznaczane jest na podstawie znajomości tłumienia ścieżki $\mathrm{PL}_i^{\mathrm{dB}}$ obliczanego ze wzoru (1) lub (2) oraz wzmocnienia anteny nadawczej $G_i^{(\mathrm{Tx})}$ i odbiorczej $G_i^{(\mathrm{Rx})}$ jako $g_i = G_i^{(\mathrm{Tx})} G_i^{(\mathrm{Rx})} 10\log_{10}\left(\mathrm{PL}_i^{\mathrm{dB}}\right)$.

2.2. Problem optymalizacyjny

Rozważany problem dotyczy maksymalizacji zasięgu sieci zakładając, że dostępna jest jedna MBS zapewniająca łączność z siecią rdzeniową i maksymalnie N UAV działających jako przekaźniki lub AP. Każdy z dronów może zostać umieszczony w jednej z dostępnych lokalizacji $L_n = (x_n, y_n)$, gdzie (x_n, y_n) to współrzędne w układzie kartezjańskim. Zakładając wymóg minimalnej przepustowości łącza dosyłowego $C_{\min}$ problem optymalizacyjny można przedstawić jako:

$$\max_{\{L_1, L_2, \ldots, L_n\}} |A|, n \le N, \tag{5}$$

gdzie $A = \{L_k, \forall k: Q_k = 1\}$ to zbiór wszystkich możliwych lokalizacji punktu dostępowego AP, dla których istnieje łącze dosyłowe od stacji MBS. Q_k to wskaźnik dostępności łącza dla lokalizacji L_k zdefiniowany jako:

$$Q_k = \prod_{i=1}^{n} q_i, n \le N, \tag{6}$$

gdzie q_i oznacza dostępność i-tego przeskoku łącza dosyłowego określoną następująco:

$$q_i = \begin{cases} 1 & \text{jeśli } C_i \ge C_{\min} \\ 0 & \text{w przeciwnym razie} \end{cases}. \tag{7}$$

Wartość C_i dla i-tego przeskoku można w praktyce zastąpić wymaganą wartością $\mathrm{SNR}_{\min}$.

W przypadku użycia RIS, takie połączenie jest uznawane za pojedynczy przeskok, gdyż nie wymaga zastosowania przetwarzania sygnału w paśmie podstawowym w RIS, a wpływ dwóch ścieżek jest uwzględniany w (2).

3. PROPONOWANE ROZWIĄZANIE I WYNIKI SYMULACYJNE

3.1. Proponowane rozwiązanie dla RIS

W celu znalezienia optymalnej lokalizacji DRS dla realizacji łącza dosyłowego należy rozważyć dla każdego przeskoku, czy między nadajnikiem i odbiornikiem dostępne jest połączenie LoS o określonej przepustowości. W tym celu można zastosować algorytm opisany w [7], gdzie zastosowano algorytm Lee dla określenia grafu widoczności $G(V, E)$, zawierającego zbiór wszystkich krawędzi budynków (E) i narożnikowych wierzchołków (V), dla których wyznaczana jest widoczność. Dla otrzymanego zbioru V_g zawierającego krawędzie między widocznymi wierzchołkami zastosowano następnie algorytm Dijkstry w celu wyboru najkrótszej ścieżki. W tej pracy rozwinięto algorytm uwzględniając dostępność nowych ścieżek związanych z użyciem RIS umieszczonych na ścianach wybranych budynków. W tym przypadku, dla algorytmu Lee RIS jest traktowany jako wierzchołek w celu znalezienia widoczności, ale nie jest uwzględniany w określaniu liczby przeskoków. Następnie znajdywana jest najkrótsza ścieżka z uwzględnieniem następującej definicji kosztu przeskoku dla algorytmu Dijkstry:

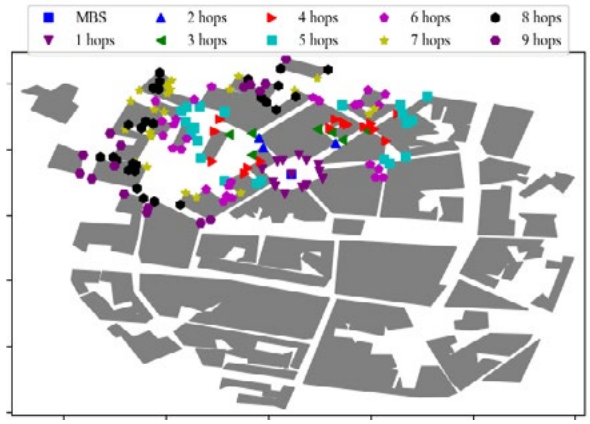
$$D_j = \begin{cases} \mathrm{PL}_{\mathrm{direct}}{}^{\mathrm{dB}} + P \text{ dla łącza bezp. i } C_i > C_{\min} \\ \mathrm{PL}_{\mathrm{RIS}}{}^{\mathrm{dB}} + P \text{ dla łącza z RIS i } C_i > C_{\min} \\ \infty \text{ w przeciwnym razie} \end{cases} \tag{8}$$

gdzie P jest stałym współczynnikiem kary z każdy kolejny przeskok, związanym z koniecznością utraty energii na przetwarzanie sygnału w przekaźniku.

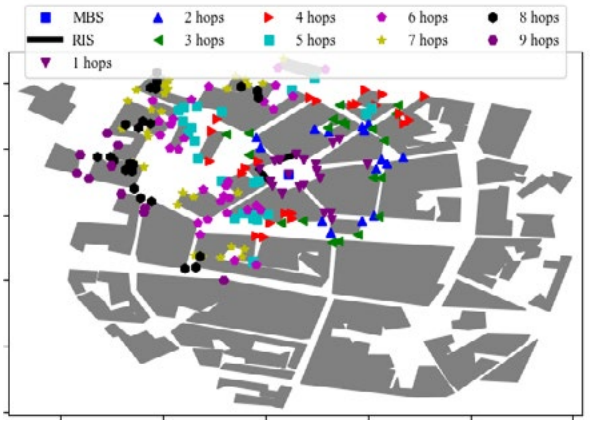
3.2. Badania symulacyjne

W celu zbadania wpływu RIS na dostępność łącza dosyłowego przeprowadzono badania symulacyjne dla systemu w dwóch konfiguracjach: bez dostępności RIS i z dwoma matrycami RIS umieszczonymi w centralnej lokalizacji obszaru (na rynku), jak przedstawiono na rys. 3. Pozostałe parametry badanego systemu dobrano zgodnie z wartościami podanymi w Tabeli 1.

Na rys. 2 przedstawiono wyniki badań dla przypadku bez użycia RIS, pokazujące liczbę wymaganych przeskoków dla wybranych lokalizacji punktu dostępowego AP (liczba przeskoków oznaczona kolorem i kształtem markera). W tym przypadku założono $\mathrm{SNR}_{\min} = 41$ dB w celu określenia dostępności pojedynczego przeskoku. Wyniki te można porównać z rezultatami otrzymanymi dla przypadku z użyciem R=2 RIS, przedstawionymi na rys. 3. Porównanie to wyraźnie wskazuje na większy zasięg systemu korzystającego z RIS. Dodatkowo, punkty dostępne również w konfiguracji tylko z wykorzystaniem dronów, w systemie z RIS można osiągnąć stosując mniejszą liczbę DRS.



Rys. 2. Ilustracja niezbędnej liczby przeskoków dla łącza dosyłowego bez zastosowania RIS dla wybranych AP



Rys. 3. Ilustracja niezbędnej liczby przeskoków dla łącza dosyłowego z zastosowaniem RIS dla wybranych AP

Tabela 1: Parametry symulacji

Parametr	Wartość
Moc nadawcza $P^{(Tx)}$	100 mW
Zysk RIS g_{bf}	15 dB
Moc szumu σ^2	-131 dBm
Efektywność łącza η	0.82
Efektywne pasmo $B^{(eff)}$	18.72 MHz
Wymagany SNR: SNR_{min} (odpowiadający C_{min})	41dB (55 Mbps)

Kolejne wyniki, przedstawione na rys. 4, pokazują maksymalną możliwą przepustowość łącza dosyłowego w zależności od lokalizacji na mapie korzystając z tzw. mapy cieplnej (ang. *heatmap*). Wartości przepustowości przedstawione na rys. 4 wyznaczono określając najlepszą ścieżkę do wybranego punktu zgodnie z zaproponowaną metodą, a następnie wybierając przepustowość najsłabszego przeskoku łącza dosyłowego. Można zauważyć, że zastosowanie większej liczby RIS w ogólności poprawia przepustowość, jednak w niektórych przypadkach może skutkować jej obniżeniem gdy umożliwia zastosowanie mniejszej liczby DRS. W tym przypadku zysk z zastosowania RIS skutkuje zmniejszeniem zużycia energii.

4. PODSUMOWANIE

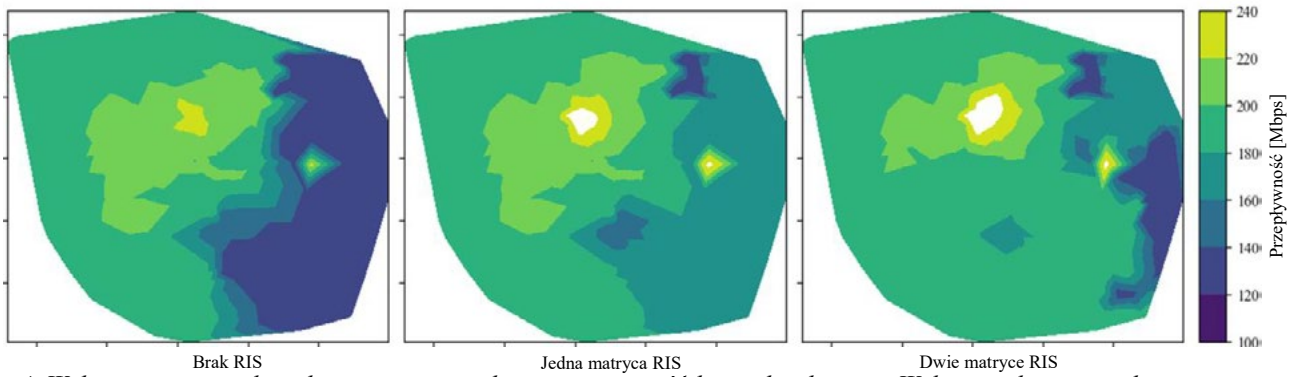
W niniejszej pracy rozważono problem realizacji bezprzewodowego łącza dosyłowego z wieloma przeskokami między stacją bazową makro i punktem dostępowym, wykorzystując do tego celu drony-przekaźniki i matryce RIS. Zaproponowano problem optymalizacyjny polegający na doborze liczby i lokalizacji dronów w celu zapewnienia odpowiedniej przepustowości łącza, dążąc do zmniejszenia liczby przekaźników lub do maksymalizacji zasięgu sieci. Przeprowadzone badania pokazały, że użycie RIS pozwala wyraźnie zwiększyć zasięg sieci lub zmniejszyć liczbę niezbędnych przekaźników do ustanowienia łącza dosyłowego. Niniejsza praca będzie dalej rozwijana uwzględniając zagadnienia maksymalizacji efektywności energetycznej i minimalizacji opóźnienia transmisji. Dodatkowo pod uwagę można też wziąć możliwość zainstalowania RIS na wybranych dronach zamiast tradycyjnych urządzeń nadawczo-odbiorczych.

PODZIĘKOWANIE

Niniejsza praca została wykonana w ramach projektu badawczego OPUS nr. 2021/43/B/ST7/01365 finansowanego przez Narodowe Centrum Nauki w Polsce.

LITERATURA

[1] Jaber, M., i in. 2016. ”5G backhaul challenges and emerging research directions: A survey.” *IEEE Access* 4 (2016): 1743-1766
[2] Challita, U., Saad, W. 2017. “Network formation in the sky: unmanned aerial vehicles for multi-hop wireless backhauling,” *IEEE Globecom*.
[3] Zhang, X.,Wang, J., Poor, H.V. 2022. “Joint Optimization of IRS and UAV-Trajectory: For Supporting Statistical Delay and Error-Rate Bounded QoS Over mURLLC-Driven 6G Mobile Wireless Networks Using FBC”. IEEE Veh. Technol. Mag., 17.
[4] Liu, Y., i in.. 2021 "Reconfigurable Intelligent Surfaces: Principles and Opportunities," *IEEE Communications Surveys & Tutorials*, v. 23, no. 3.
[5] Jeon, H. -B., i in. 2021 "RIS-assisted Aerial Backhaul System for UAV-BSs: An Energy-efficiency Perspective," *IEEE Globecom*.
[6] Mohamed, E. M., i in. 2023. "Distribution of Multi MmWave UAV Mounted RIS Using Budget Constraint Multi-Player MAB," *Electronics*, vol .12.
[7] Janji, S., i in. 2022. "On the Placement and Sustainability of Drone FSO Backhaul Relays," *IEEE Wireless Communications Letters*, vol. 11, no. 8.
[8] Rappaport, T. S., i in. 2013 "Broadband Millimeter-Wave Propagation Measurements and Models Using Adaptive-Beam Antennas for Outdoor Urban Cellular Communications," *IEEE Transactions on Antennas and Propagation*, vol. 61, no. 4.
[9] Abdel-Malek, M. A., i in. 2019. “UAV positioning for out-of-band integrated access and backhaul millimeter wave network,” *Physical Comm.*, vol. 35.
[10] Kishk, M. A., Alouini, M. -S.. 2021. "Exploiting Randomly Located Blockages for Large-Scale Deployment of Intelligent Surfaces," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 4.
[11] Mogensen, P. i in. 2007. "LTE Capacity Compared to the Shannon Bound," *IEEE 65th Vehicular Technology Conference - VTC2007-Spring*.



Rys. 4. Wykresy mapy cieplnej ilustrujące osiągalną przepustowość łącza dosyłowego. Wykres po lewej przedstawia sytuację bez RIS, środkowy - z wykorzystaniem jednej matrycy RIS, a prawy - z zastosowaniem dwóch RIS

Implementacja sieci rdzeniowej 5G w środowisku chmurowym

Cloud-enabled deployment of 5G core network

Mateusz Zięba[1]; Piotr Boryło[1];

[1] AGH Akademia Górniczo-Hutnicza,
mateuszzieba@student.agh.edu.pl,
piotr.borylo@agh.edu.pl

DOI: 10.15199/59.2023.4.16

Streszczenie: Oprogramowanie stanowi nieodłączny element systemów telekomunikacyjnych. Jest to widoczne w sieciach 5G, które definiują funkcje sieciowe jako procesy programistyczne, z możliwością wirtualizacji. Kolejnym krokiem jest uniezależnienie wdrożeń od sprzętu oraz nadzorcy wirtualizacji. W pracy, przedstawiamy wdrożenie sieci rdzeniowej 5G w środowisku chmurowym. Przeprowadziliśmy także walidację oraz zaproponowaliśmy sposób rozwiązania problemu. Wdrożenie rozszerzyliśmy o funkcję sieciową na potrzeby analityki.

Abstract: The ongoing process of network *softwareization* is clearly visible for the mobile networks. The 5G standard defines core network functions as separate processes that can be easily virtualized. The next step is to make deployments cloud-enabled, e.g., independent of the hardware and hypervisor. In this work, we present a Proof-of-Concept of the 5G network cloud-based deployment. We perform the troubleshooting and propose solution for the encountered problem. We also extended the existing implementation with an analytics function.

Słowa kluczowe: sieci 5G, wdrożenia w chmurze obliczeniowej, wirtualizacja, sieci logiczne (slicing)

Keywords: 5G network, cloud-based depolyments, virtualization, network slicing

1. Introduction

The deployment of 5G mobile core networks has become crucial due to the increasing demand for fast and reliable connectivity. Cloud infrastructure has emerged as a promising solution to enhance the scalability and flexibility of such a deployment. Virtualization, enables operators to effectively meet user demands and reduce costs. Simultaneously, collaboration between network service providers and device manufacturers has led to the increased availability of 5G-enabled devices.

Cloud services have gained popularity, with businesses of all sizes offering tailored cloud solutions to meet customer needs. Global trends emphasize the development of cloud-based applications, data hosting, and even complete network deployments as the new standard. Mobile network vendors and operators recognize the advantages of employing cloud solutions in mobile telecommunications, thus establishing a strong connection between 5G networks and the cloud.

The contribution of our work can be summarized as follows. Firstly, we present a Proof-of-Concept (PoC) of the 5G network deployment in the virtualized environment and with the support for slicing (Section 2). Secondly, we performed the PoC verification process that revealed the error-prone scenario (Section 3). Thirdly, proposed the original solution for the problem (Section 3.2). Finally, we extended the initial deployment with an additional module enabling analytics functions (Section 4). The aim of our work is to facilitate cloud-enabled deployments of 5G network for the research purposes. Supplementary materials are published in public repository [6].

1.1 Virtualization in 5G core networks

As the evolution in mobile networks progressed, one aspect remained intact: the reliance on 'black- box' solutions in terms of mobile network components. The vendors provided integrated hardware with dedicated software that would properly interact only with other proprietary equipment. To overcome the limitations, 5G networks utilize virtualization of their core components and use commercial off-the-shelf (COTS) hardware. Virtualization allows for additional flexibility, as Network Functions (NFs) can be dynamically scaled in or out according to the network state. Additionally, it reduces time-to-market [4] of new services, since the adaptation time of network architecture is reduced. In principle, virtualization significantly impacts both operational (OPEX) and capital (CAPEX) expenses. CAPEX of deploying 5G core networks is greatly reduced as NFs no longer require to run on dedicated, costly hardware. OPEX of maintaining such aggregation of virtualized resources and adjusting them to provide novel services is also reduced.

5G networks have been designed to support deployment that uses the concept of Network Function Virtualization (NFV) [1]. This virtualization framework describes the network node functions as a software-entities that can run independently of the hardware. There are three main domains of NFV framework: (1) Virtual Network Function (VNF), which is a software implementation of a network component, capable of running on various platforms; (2) Network Function Virtualization Infrastructure (NFVI) that denotes the COTS hardware that VNFs are deployed on; (3) NFV management and orchestration (NFV-MANO) framework for administrating VNFs and NFVI.

There are two ways of deploying NFs. The first one assumes that NFs are deployed using containerized network services [2]. A container is a lightweight, standalone unit of software, encapsulated with the runtime environment and all required dependencies. Implementing the core network components as containerized functions in clustered worker nodes enables a straightforward vertical scaling, increased reliability and swift fault recovery [3]. The second way to virtualize network functions is to use virtual machines. In this deployment type, a NF runs as a process on the guest operating system inside a VM. A VM-based NF requires a commodity hardware server and an underlying operating system with various dependencies. In general, VM-VNF requires more disk space as it includes an operating system necessary to host the NF process. The size of the NF implemented as a VM-VNF affects the time required for deployment and migration. Additionally, hardware virtualization consumes a significant portion of the physical resources, such as the CPU. Despite the disadvantages of the virtual machine approach, VNFs have been successfully deployed on top of VMs in many commercial deployments.

1.2 Network slicing

As 5G network supports diversified services, it must meet a wide range of requirements for capacity, latency, isolation or resilience. These requirements are often impossible to be achieved simultaneously. Thus, the mechanism responsible for network isolation, called slicing, has been introduced.

A slice is a logical network with components customized to meet the needs of a specific service. It uses a dedicated core network architecture and provides access to services located in external data networks (DNs). A DN can provide operator services, internet access, or third-party services. [5]

User Equipment (UE) might request access to a specific slice, based on the Single Network Slice Selection Assistance Information (S-NSSAI). S-NSSAI is composed of SST and SD fields. Upon registration, UEs configured with different S-NSSAIs should access the corresponding network slice and establish PDU session accordingly. It enables slicing in the initial registration and service request.

Initial registration with PDU session establishment are key operations to be supported by 5G network. The joint procedures can be used to determine that a deployed 5G network operates properly. Therefore, it is our starting point to verify the proposed PoC.

2. The Proof-of-Concept

In this section, we present the PoC of the cloud-enabled 5GCN deployment supporting network slicing and virtualization of NFs. The 5GCN network functions are deployed as separate processes on virtual machines. NFs of a 5G core network are hosted on virtualized nodes with the use of type-2 hypervisor. We utilized VirtualBox (VBox) hypervisor but resulting images of VMs are easily portable and can be integrated with a selected cloud infrastructure.

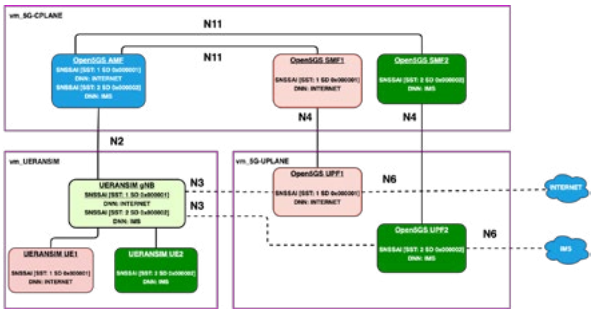


Figure 1: Virtualized 5G core network with slice-dedicated NFs.

Essentially, the platform for hosting 5G core and access networks consists of three VMs. Each virtual machine is equipped with a northbound interface used for remote access by the host. The actual communication related to 5GCN is handled via data interfaces and the traffic is routed via an internal virtual router. Such a networking architecture is typical for cloud environments. Open5GS is chosen as the 5th generation mobile core network framework suitable for the proposed architecture. The selection has been motivated by the broad and precise documentation, the number of services provided by NFs and the vast community around the project. Since an emulated access network, together with UE, are required for a proper verification of 5GCN, an additional virtualized node is deployed for 5G Access Network (5G-AN). UERANSIM is an open source framework delivering 5G-AN functionalities. It provides a software implementation of gNodeB (gNB), which is a base station for 5th generation AN, and a 5G-enabled UE. In the deployed 5G network, the Service Based Architecture (SBA) communication via SBI interfaces and N2/N1/N4 link message exchange is routed over private address space.

To sum up, the architecture is shown in Figure 1 and has been carefully designed to ensure repeatability and facilitate cloud migration. Thus, the 5GCN and access network can be fully deployed in a public cloud.

2.1 Setup parameters and networking

The virtual machines are assigned with the following roles:

- **vm_5GC-CP** hosts the control plane NFs,
- **vm_5GC-UPF** hosts 5GC User Plane Functions,
- **vm_5G-UERANSIM** hosts the 5G access network together with UEs.

Table 1: VBox guest platforms parameters.

VM name	RAM	disk space
vm_5GC-CP	2048 MB	6 GB
vm_5GC-UPF	1024 MB	6 GB
vm_5G-UERANSIM	1024 MB	6 GB

The configuration of the physical parameters of the deployed VMs is presented in the Table 1. To accommodate more NFs deployed on **vm_5GC-CP** compared to the other VMs, its physical resources (i.e., RAM) have been doubled. Each VM has been equipped with the 64-bit version of *Debian 11 Bullseye* OS.

When a VM generates any outbound traffic, the NAT adapter of the hypervisor's NAT engine extracts the TCP/IP data and forwards it through the host operating system to a remote network (e.g., the internet). Each VM is equipped with three network adapters.

The first one, i.e. *enp0s3* for Debian distributions, is configured as NAT adapter to enable access to the internet from within the VM. The NAT interface is used as the default gateway device. This address is bounded to the loopback interface of the host OS. The IP address of the NAT interface of the guest VM is acquired from the hypervisor's DHCP engine and resides within a NAT address pool (10.0.2.0/24). The remaining two interfaces of each VM are configured as *host-only networking* type. Such a networking environment is easily reproducible in the cloud.

2.2 Slicing configuration

In the deployed PoC, 5GCN is logically divided into two slices: *internet* [SST:1, SD: 0x000001] and *ims* [SST:2, SD: 0x000002]. Both are dedicated to provisioning a data service. The configuration presented below is critical to deploy slice-enabled and virtualized 5GCN in cloud environment. All the configuration files are public [6].

As presented in Figure 1, the slices are assigned with a dedicated SMF function in Control Plane (CP), and a UPF in Data Plane (DP). The slices can be distinguished by referring to the respective colour (*pink* or *green*) of the NF blocks and the S-NSSAI mentioned within the NFs. The remaining NFs of the 5G CP plane are shared among the two slices, these are: NRF, UDR, UDM, BSF, PCF, AUSF, NSSF and AMF.

SMF1, deployed on *vm_5GC-CP*, binds its PFCP interface (the N4 interface in reference-point architecture) to *enp0s9* physical interface, with an IP address of 192.168.58.111. The same line is present in UPF1 configuration to ensure a N4 link is established and PFCP peering has been achieved. The UPF1 configuration file has been edited in order to alter the default PFCP address, which now points to the same network as PFCP address of SMF1. The PFCP address is bound to the *enp0s9* physical interface on *vm_5GC-UPF* VM, and equals 192.168.58.121. The GTP-U interface, i.e. the N3 reference point for data transfer, also points to 192.168.58.121 address. The *subnet* values reflect the SMF1 configured IP address pool and DNN. The *dev* value denotes the tunnel interface name, which is used for data tunneling via N6 interface to DN, i.e. public internet.

The configuration of SMF2 and UPF2 for slice [SST:2 SD:0x000002] support is similar to the SMF1/UPF1. The changes however reflect the fact that both SMF1/SMF2 and UPF1/UPF2 are based on the same VM: *vm_5GC-CP* and *vm_5GC-UPF* respectively. Therefore, the IP addresses of PFCP and GTP-U interfaces, although bound to the same interfaces as addresses SMF1/UPF1, are changed. For SMF2, PFCP: 192.168.58.112 and GTP-U: 192.168.58.112. For UPF2, PFCP: 192.168.58.122 and GTP-U: 192.168.58.122.

Due to the fact that NFs in the 5GCN CP are deployed on one virtual machine, the SMF2 SBI interface cannot point to the same IP address as the loopback interface of the *vm_5GC-CP*. The SMF2 SBI address has been changed to 127.0.0.24 (port remains intact). This change is later reflected upon NF registration in the NRF, thus the SMF2 services are accessible via the altered SBI.

Furthermore, the *freeDiameter* configuration file has been swapped to indicate the distinction between SMF1 and SMF2 in terms of SBI addresses (*ListenOn = "127.0.0.24"*).

3. Verification of the PoC

The scenario to be realized using the mentioned 5G network architecture is as follows. Two UEs (*UE1, UE2*) send a registration request via Non Access Stratum (NAS) message to a shared AMF. As the UEs are authorized, the separation related to slicing happens. The AMF performs SMF selection procedure and finds a SMF suitable for the particular S-NSSAI and data network. If the selection is positive, AMF requests PDU session creation from the SMF. As a result, the PDU session for a given UE should be created in a respective logical network. That is, UE1 PDU session request to *internet* DN should be managed by SMF1, and the User Plane tunnel for its data should be anchored in UPF1. The exact procedure should be repeated for the UE2 PDU session establishment request. Supplementaty figures and listings are published in [6].

The verification of the virtualized 5GCN consists of two subparts: Initial Registration and UE triggered service request. However, since the UEs are configured to initiate the PDU session request immediately after a 5GN access, both procedures are combined. Although the registration to the network was successful, the session establishment failed.

3.1 Troubleshooting process

After multiple attempts to establish a PDU session, the UE1 fails to receive any response to its *InitialUEMessage* and switches to the IDLE state. The AMF logs allowed tracing the cause of this issue in the implemented 5GCN. Namely, the AMF's inability to connect to a HTTP/2 server at the 127.0.0.24 address.

This address belongs to the SBI interface of SMF2. The AMF requested session establishment in an invalid SMF, since SMF2 is located in a different slice, the UE1 PDU session establishment request should not have been forwarded to this specific network function. During further investigation, it was revealed that the SMF2 has received a *Nsmf_PDUSession_CreateSMContext* request for a DN with unsupported slice combination: [SST: 1, SD:0x000001].

Before issuing the faulty request, the AMF performs an internal SMF selection procedure. This procedure, may be skipped if the AMF has local access to SMF information. This explains the lack of the actual SMF selection procedure, as both SMF1 and SMF2 are located within one NSI and are registered with a common NRF. The AMF obtains SMF information through a notification sent by the NRF to the */nnrf-nfm/v1/nf-status-notify* endpoint and stores it until its deregistration.

We adopted a reverse engineering approach to verify why AMF fails to anchor the PDU Session in the adequate SMF. The cause of the faulty SMF selection may be related to the *nf_instance* itself. Since SMF2 was called after SMF1, it may have been chosen as the desired node for requesting the session establishment. To address this issue, the execution order of the Session Management Functions was reversed. Contrary to the expectations, a reversed execution of SMFs did not solve the issue regarding SMF selection mismatch. AMF continued to is-

sue *Nsmf_PDUSession_CreateSM- Context* Requests for UE1 PDU session establishment to an invalid SMF2.

Unfortunately, the documentation of Open5GS does not provide the possible fix for the SMF selection issue. The abnormal behavior might be related to the case when multiple S-NSSAIs are associated with one NSI, however this has not been proven.

3.2 Proposed solution

To propose the effective solution, we experimented with procedures related to the UE-triggered service request and NF deregistering. Based on that, we proposed a solution to the presented problem that enables slicing in a cloud-enabled and virtualization-based 5G deployments.

The procedure is as follows: (1) ensure that NFs have registered to the NRF, with an exclusion of SMF2; (2) trigger a registration with PDU Session establishment for the UE1; (3) verify that the PDU Session has been anchored in UPF1 and the NF is capable of tunneling UE traffic to an outbound DN; (4) utilize the UERANSIM provided CLI or UE1 and trigger the UE-initiated PDU session request; (5) invoke the NF deregister procedure of SMF1 by stopping its process on vm_5GC-CP VM; (6) call SMF2 executable and trigger UE2 registration; (7) verify that PDU Session for UE2 has been anchored in the correct slice; (8) deploy SMF1 NF; (9) verify that PDU Session for UE1 has been reestablished (i.e. AMF issued the request to SMF1 instead of SMF2), avoiding the crash scenario.

The UE1 *Nsmf_PDUSession_CreateSM-Context* Request has been successfully issued to the SMF1, which registers the PDU session in Policy Control Function (PCF). It proves a positive workflow of a PDU Session establishment procedure and correct UPF was selected for the DN requested by UE1. Also, the UE2 request for a PDU service has been correctly handled. Since AMF session cache had been cleared upon being notified of SMF1 deregistration by the NRF, the internal SMF selection process has returned the only available instance of SMF2. The PDU Session has been successfully established by UE2 in the requested slice.

4. Extension for analytics of 5G deployment

Common open source implementations of the 5G Core Network includes a sufficient number of NFs to meet the specification of Release 16 [1]. However, 5G cloud-based deployments in virtualized environment with the support for network slicing may require supplementary functions.

Namely, mobile network operators would expect a Network Exposure Function (NEF) to provide 5G core internal data for broad analysis. Such analysis based on on-demand live data from the network would be then used to foresee undesired events, i.e. overloads, service request rejects. Summarizing slice-specific data can be useful when predicting the resources that need to be deployed to ensure uninterrupted service provisioning. For example, aggregated data that shows the total aggregated maximal bitrate for a given DN may suggest increasing the number of UPF instances in the User Plane to reduce pressure on the existing NFs. If the 5GCN CP needs to be scaled, information about the SSCs supported by UEs for a given DN can help to ensure PDU Session continuity as the generated traffic is offloaded onto newly deployed UPFs. The total bitrate per DN may also be a factor to consider when adjusting virtualized resources to the network state, enabling efficient cloud-based deployment. If the 5GCN infrastructure is at risk of being overloaded, additional network resources must be deployed. The pressure on the physical hardware should be reduced to ensure that 5G critical services can continue to be delivered.

The current implementation of the framework does not provide described functionality. Thus, we designed, proposed and implemented an interface to the Unified Data Repository function. The aim is to propose a quasi-NEF that can be easily integrated with virtualized cloud-based deployments of 5G core network. The quasi-NEF's goal is to aggregate 5G related data that would be useful for a mobile network operator or a private 5G slice tenant. The implemented function is capable of: summarizing slice-related information, listing all preemption vulnerable and preemption capable UEs.

5. Conclusions

In this work, we presented a Proof-of-Concept of the cloud-based deployment of 5G network with the use of virtualization. Our installation enables slicing, that may be challenging to be configured. Thus, to facilitate similar deployments, we provide details on our configuration. We also performed an initial verification of the PoC, found the error-prone scenario and proposed the original solution that enables similar deployments in the cloud environment. Finally, we extended the existing implementation with an additional network function, denoted as quasi-NEF. Our module extends the functionality towards analytics, being a critical issue in complex, virtualized environments. Configuration files, source codes, listings and additional figures are available in [6].

We plan to perform comprehensive performance tests on the prepared deployment, verify the migration process to the public cloud and compare the performance of local and cloud installations. It will be also beneficial to develop and automate the proposed solution, as currently it involves dynamic phasing in and out of network functions to provide the requested service. Finally, further steps may include analogous container-based deployment to be compared with current one, using virtual machines.

References

[1] 3GPP, 2020, *System architecture for the 5G System (5GS)*, TS 23.501 version 16.6.0

[2] Leonardo Bonati et al, 2020, *Open, Programmable, and Virtualized 5G Networks: State-of-the-Art and the Road Ahead*, Computer Networks, 182

[3] Luong, Duc-Hung and Thieu et al, 2018, *Cloudification and Autoscaling Orchestration for Container-Based Mobile Networks toward 5G: Experimentation, Challenges and Perspectives*, 2018 IEEE 87th Vehicular Technology Conference (VTC Spring)

[4] Sploty Sp. z o.o., 2022, *5G System Overview*

[5] Wan Lei and Anthony C. K. Soong et al, 2020, *5G System Design. An End to End Perspective*, Springer

[6] M. Zięba and P. Boryło, 2023, *Supplementary materials*, https://github.com/vvafwgsv/KRiT2023-Cloud-enabled-deployment-of-5G-core-network, (access date: 2023-05-17)

Analiza pomiaru współczynnika błędów dla obcinania przekonań i kwantyzacji w kodach polarnych

Measurement Analysis of Error Rate for Belief's Clipping and Quantization in Polar Codes

Yasir Ali[1], Wojciech Sułek[2]
[1] Politechnika Śląska, Gliwice, Yasir.Ali@polsl.pl
[2] Politechnika Śląska, Gliwice, Wojciech.Sulek@polsl.pl

DOI: 10.15199/59.2023.4.17

Streszczenie: Kody korekcyjne, do których należą kody polarne, odgrywają kluczową rolę w niezawodnej komunikacji w obecności szumu, interferencji i innych zniekształceń. W tej pracy wykonujemy badania eksperymentalne dwóch modyfikacji, a mianowicie obcinania wartości LLR(Log-Likelihood Ratio) i ich kwantyzacji, oraz prezentujemy wynikowe bitowe stopy błędów dla dekodowania SC (Successive Cancellation) kodów polarnych. We wnioskach przedstawiamy kompromis pomiędzy poziomem obcięcia wartości LLR, liczbą poziomamów kwantyzacji a skutecznością dekodowania.

Abstract: Error-correcting codes, such as polar codes, play a critical role in reliable communication in the presence of noise. In this study, we investigate the impact of two modifications, namely Log-Likelihood Ratio (LLR) clipping, and quantization, on the resulting error rates for Successive Cancellation (SC) decoding for polar codes. Our evaluation involves processing the clipped and quantized LLR vector through the decoder while examining the trade-off between clipping, quantization levels, and decoding accuracy.

Słowa kluczowe: Obcinanie LLR, kwantyzacja LLR, kompromis między wydajnością a kwantyzacją dla SCD.

Keywords: LLR clipping, LLR quantization, Performance-quantization tradeoff for SCD.

1. INTRODUCTION

Polar codes have emerged as a promising class of codes that achieve the capacity of symmetric binary-input discrete memoryless channels. Successive Cancellation (SC) decoding is a widely used decoding algorithm for polar codes due to its simplicity and low complexity. However, the performance of SC decoding can be significantly affected by the presence of excessive noise in the received Log-Likelihood Ratios (LLRs).

While significant advancements have been made in decoding polar codes over additive white Gaussian noise (AWGN) channels, many practical communication scenarios involve non-Gaussian noise sources. Examples include non-linear demapping for soft decoding in the presence of impulsive noise [1], underwater acoustic signal denoising algorithm for Gaussian/non-Gaussian impulsive noise [2], and Bi-Gaussian approximation under multiuser interference [3]. These scenarios exhibit impulsive noise, characterized by short-duration interference with large amplitudes that occur more frequently than predicted by Gaussian distributions. The presence of such impulsive noise can significantly degrade the performance of soft decision decoders, particularly when the noise parameters are incorrectly estimated.

LLR clipping involves limiting the LLR values to a defined range, effectively mitigating the impact of excessive error events. By constraining the LLRs within an acceptable magnitude, the decoding algorithms can operate on more stable LLR estimates. Several studies have shown the benefits of LLR clipping in improving the performance of different communication systems, such as, in [4] radio wave performance has been improved in MIMO communication when LLR is clipped to average, in [5] the performance of LDPC codes is enhanced by using a simple clipping approach that reduces the impact of large noise values of impulsive interference, and [6] demonstrated the decrease in computational complexity by implementing a technique known as simple adaptive LLR clipping.

Furthermore, LLR quantization is introduced to remap the continuous LLR values to a reduced set of discrete levels. This quantization enables to control the resolution of LLR values, enhancing the ability to differentiate between levels and allowing for convenient modeling of the real, constrained precision, hardware implementation. Various quantization schemes have been investigated in the literature. For example, in a recent study [7] of channel decoding, it was demonstrated that polar codes under SC decoding exhibit improved error-correction performance for low-precision decoding. Specifically, in the case of the MK code with a length (1024,512), the introduction of LLR quantization positively impacts the decoding performance. Moreover, irrespective of the polar coding, [8] showed that the use of constrained digital arithmetic can significantly reduce the memory requirements and computational complexity of decoding algorithms. It has been proposed in [9] that SC-list polar decoding with 7-bit quantization can be pruned to achieve a target BLER, making it possible to implement ultra-reliable low latency communication (URLLC) in wireless solutions with reduced memory requirements. Additionally, the authors of [10] conducted an analysis of the robustness of polar codes with a quantized decoder and concluded that polar codes are robust when subjected to quantization.

The primary objective of this study is to analyze the impact of LLR clipping and quantization on the performance of SC decoding in polar codes, by simulation experiments. We aim to assess the tradeoff between performance and quantization levels, by evaluating the Bit Error Rate (BER) under varying Signal-to-Noise Ratio (SNR) conditions. The experimental evaluation will be conducted using simulation frameworks and performance metrics widely adopted in the field of polar code research.

By comparing the performance of the modified SC decoding algorithm with the conventional SC decoding, we can gain insight into the effectiveness of the proposed mo-

difications and their impact on the decoding performance. This research will contribute to the development of improved decoding algorithms for polar codes, addressing the challenges posed by noisy channels and efficient hardware implementations of decoders.

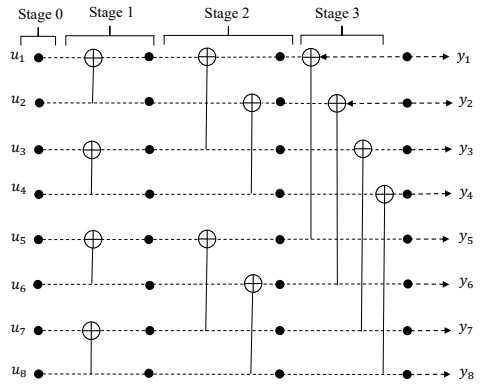


Fig. 1: *A polar code $\mathcal{P}(8,4)$ tree traversing diagram.*

2. PRELIMINARIES

2.1 Polar Codes

A polar code $\mathcal{P}(N,K)$ with code length N and information length K can be constructed by assigning K reliable bits $u_{\mathcal{A}}$ and $N-K$ unreliable bits $u_{\mathcal{A}^c}$ in the code vector u, where $\mathcal{A}$ represents the set of information bits and $\mathcal{A}^c$ represents the set of frozen bits. The predefined value is used to set $N-K$ frozen bits. For example, a polar code $\mathcal{P}(8,4)$ tree traversing diagram is shown in Fig. 1. The codeword is obtained by multiplying the source bits u_1^N with the generator matrix G_N, which is expressed as follows:

$$x_1^N = u_1^N G_N = u_1^N F_2^{\otimes n} \tag{1}$$

where, $n = \log_2 N$, and $F_2^{\otimes n}$ is the n-th Kronecker product of the polarization matrix,

$$F_2 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

2.2 Successive Cancellation Decoding

Let $\alpha_{0,j}$ be the LLR of the channel output y_j, and let $\alpha_{i,j}$ denote the j-th internal LLR in the i-th decoding stage, where $i = 1,2,\ldots,n = \log_2 N$, and $j = 1,2,\ldots,N$. Generally, the LLR $\alpha_{i,j}$ is computed as the natural logarithm of the ratio between two conditional probabilities. It measures the relative likelihood of observing the received sequence and estimated values under the assumption that the i-th bit is 0 compared to the assumption that it is 1. The LLR is a crucial parameter used in the decoding process of polar codes to make decisions about the transmitted bits based on received observations. The LLR $\alpha_{i,j}$, is defined in [10] as:

$$\alpha_{i,j} = \ln \frac{P\left(\boldsymbol{y}, \hat{u}_1, \hat{u}_2, \ldots, \hat{u}_{j-1} \mid u_j = 0\right)}{P\left(\boldsymbol{y}, \hat{u}_1, \hat{u}_2, \ldots, \hat{u}_{j-1} \mid u_j = 1\right)} \tag{2}$$

where, $\alpha_{i,j}$ refers to the LLR value for the j-th bit in the i-th level of the polar code;
$P(\boldsymbol{y}, \hat{u}_1, \hat{u}_2, \ldots, \hat{u}_{j-1} \mid u_j = 0)$ represents the conditional probability of receiving the sequence of observations $\boldsymbol{y}$, along with the estimated values $\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_{j-1}$, given that the j-th bit (u_j) in the polar code is set to 0;
$P(\boldsymbol{y}, \hat{u}_1, \hat{u}_2, \ldots, \hat{u}_{j-1} \mid u_j = 1)$ represents the conditional probability of receiving the same sequence of observations and estimated values, but with the j-th bit (u_j) set to 1.

To compute the internal LLRs for SC decoding in a recursive and low-complexity manner, the min-sum operation is known to be effective, as given below,

$$\begin{aligned} \alpha_{i,j} =& \operatorname{sign}\left(\alpha_{i-1,j}\right) \cdot \operatorname{sign}\left(\alpha_{i,j+N/2^i}\right) \\ & \times \min\left\{\left|\alpha_{i-1,j}\right|, \left|\alpha_{i-1,j+N/2^i}\right|\right\} \end{aligned} \tag{3}$$

$$\alpha_{i,j+N/2^i} = (-1)^{\hat{u}_{i,j}} \cdot \alpha_{i-1,j} + \alpha_{i-1,j+N/2^i}, \tag{4}$$

where $\hat{u}_{i,j}$ is the hard decision result for the internal bit $u_{i,j}$. The functions (3) and (4) are commonly referred to as the f and g functions in polar decoding. To decode a bit, the estimation is obtained based on its LLR and whether it is a frozen bit, using the following rule:

$$\hat{u}_j = \begin{cases} 0, & \text{if } j \in \mathcal{A}^c \\ 0, & \text{if } \alpha_{i,j} < 0 \\ 1, & \text{otherwise} \end{cases}. \tag{5}$$

This equation serves as a decision rule to estimate the value of $\hat{u}_j$ based on the LLR value α and the index i, taking into account the decoding process and the likelihood of the bit being 0 or 1.

3. METHODOLOGY

In this work, we investigate the system model illustrated in Fig. 2. The message vector $u_{\mathcal{A}}$ undergoes non-systematic polar encoding with beta-expansion [10] to generate the codeword x_1^N. BPSK modulation is then employed to modulate the codeword, resulting in the modulated symbols vector s_1^N. The modulated symbols are transmitted over the AWGN channel with noise variance σ^2.

To transform the received soft value y into the LLR soft value r at the receiver, an LLR converter is used, given by:

$$r = \mathrm{LLR}(y) = \ln \frac{p(y \mid u=0)}{p(y \mid u=1)} = \sigma^2 y \tag{6}$$

Note that the conditional distributions $p(r \mid x=+1)$ and $p(r \mid x=-1)$ are Gaussian distributions with mean values $\pm 2/\sigma^2$, respectively, and variance $4/\sigma^2$. Therefore, the unconditioned distribution of LLRs is a bimodal Gaussian distribution.

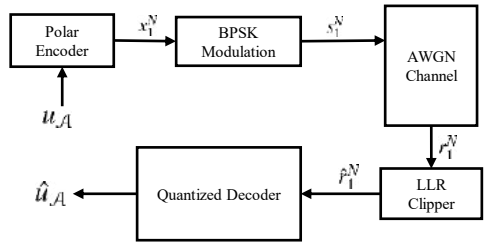


Fig. 2: *Block diagram of the system model.*

3.1 LLR Clipper

The basic function of LLR clipper is to cut the excessive noise and, secondly, to enable constrained precision representation, as depicted in Fig. 3. By limiting the LLR values to a defined range of $-r_{max} < r < +r_{max}$, the LLR clipper ensures that the LLRs remain within an acceptable magnitude to reduce noise and preserve essential

information. Mathematically, the clipped LLR $r^{'}$ can be represented as follows:

$$r^{'} = \begin{cases} r, & \text{if} \quad -r_{\max} < r < +r_{\max} \\ -r_{\max}, & \text{if} \quad r < -r_{\max}, \\ +r_{\max}, & \text{if} \quad r > +r_{\max}. \end{cases} \tag{7}$$

where r represents the original noise-aided LLR value, $+r_{max}$ is the maximum positive threshold, $-r_{max}$ is the maximum negative threshold. The function $r^{'}$ ensures that the LLR is constrained within the defined range. The clipped symbol vector $\hat{r}_1^N$ is fed to the quantized decoder, which provides the estimation of the message vector $\hat{u}_{\mathcal{A}}$, based on the chosen quantization scale $q_r = 2, 4, 8, 16, 32$. The application of the LLR clipper enables a controlled reduction of noise, allowing subsequent decoding algorithms to operate on LLRs with improved reliability.

The LLR clipper finds significance in various communication systems, particularly in other scenarios, where channels are prone to high noise levels or have limited dynamic range for radio waves in MIMO communication [4]. By suppressing excessively noisy LLRs, the LLR clipper improves the error correction capability of the decoding algorithms in LDPC decoding, leading to improved overall system performance [5]. However, the application of LLR clipping in the context of polar decoding has not yet been explored. Therefore, our approach focuses on a balance between noise reduction and preserving critical information, ensuring robust and accurate decoding outcomes of the polar decoding.

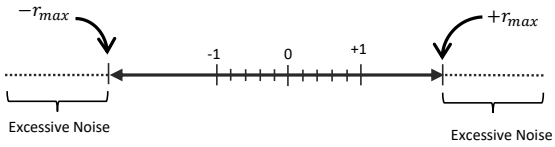


Fig. 3: *LLR clipper for* r_{max}.

3.2 Quantized SC Decoder

By applying quantization, the LLR values from the LLR clipper are uniformly transformed into distinct quantization levels. The LLR quantizer typically employs quantization levels of $2 * 2^n$, such as 2, 4, 8, 16, or 32 levels, both on the positive and negative sides of the scale, allowing for a fine-grained resolution of LLR values. Mathematically, the LLR quantizer remaps the LLR values as follows:

$$L_q = Q(r^{'}) = floor\left(\frac{r^{'}}{r_{\max}} \times q_s\right), \tag{8}$$

where $r^{'}$ represents the output of the LLR clipper, $r_{\max}$ represents the maximum threshold of the LLR clipper, q_s is the quantization levels in terms of $2 * 2^n$ as $r^{'}$ has both negative and positive values, $Q()$ denotes the quantization operation, and L_q is the quantized LLR value for $r^{'}$. The quantization step size determines the granularity of the LLR quantization, allowing for a controlled tradeoff between computational complexity and decoding accuracy. Smaller step sizes result in finer resolution but require additional computational precision, which entails more hardware resources. On the other hand, larger step sizes reduce computational overhead but sacrifice resolution. The choice of quantization levels and step size allows for customization to meet specific requirements and tradeoffs.

By remapping the LLR values to discrete levels, the LLR quantizer facilitates various decision-making processes in decoding. Soft decisions involve utilizing the quantized LLR values directly, while hard decisions involve thresholding the quantized LLR values to obtain binary decisions.

The LLR quantizer enables efficient hardware implementation by reducing memory requirements and computational complexity. Additionally, it is a known effect that the integration of the LLR quantizer in error-correcting codes enhances the ability to eliminate channel impairments, such as noise and interference. By reducing the LLR values to a discrete set of levels, the quantization process improves the resilience of the decoding algorithm against quantization errors and quantization-induced performance degradation.

4. SIMULATION RESULTS

In this section, we will present the simulation results of the proposed modifications in the SC decoding algorithm for polar codes with varying lengths N and rates $R = K/N$.

The numbers of quantization levels q_s explored in this simulation include the 2, 4, 8, 16, and 32 levels, symmetric with respect to 0. To assess the impact of proposed LLR clipping and quantization, simulations were conducted for an SC decoding of polar codes, by varying the E_b/N_0 (Energy per Bit to Noise Power Spectral Density ratio) over a range of values. We transmitted 100 encoded blocks of lengths $N = 1024$, and $N = 512$, $r_{\max} = 3$ while keeping E_b/N_0 constant for each curve. We observe the BER response with variation in quantization levels q_s.

We conducted simulations to analyze the performance of the system using different rates, specifically $R = 5/7$ and $R = 1/3$. For the case where the number of information bits K is greater than the frozen bits, denoted as $\mathcal{P}(1024, 732)$ and $\mathcal{P}(512, 362)$, as well as for the case where the number of information bits is lower than the frozen bits, denoted as $\mathcal{P}(1024, 342)$ and $\mathcal{P}(512, 171)$, and obtained BER plots. The simulation results for these groups can be observed in Fig. 4 and Fig. 5, respectively. In the plots, the solid lines correspond to the rate $R = 5/7$, while the dotted lines represent the rate $R = 1/3$.

It can be observed that the BER curves exhibit different responses against each step of E_b/N_0, and produce a separate curve along with the variation of quantization levels. As the number of quantization levels increases, the decoding accuracy tends to improve due to a finer representation of the LLR information. However, the performance gain diminishes as the number of levels becomes excessive. By analyzing the results obtained from both LLR clipping and quantization, a tradeoff analysis can be conducted to determine the optimal combination of these modifications for polar code decoding. The tradeoff is primarily concerned with achieving a desirable level of noise reduction while minimizing the loss of valuable LLR information and controlling computational complexity. For this purpose, observing the BER curves of both scenarios, it is evident that beyond a quantization level of $q_s = 16$, the curves start to flatten out, indicating negligible gain with increasing quantization levels. Therefore the balanced tradeoff for the error rate and quantization could be considered at $q_s = 16$. The discontinuation of values at certain points in the graph occurs when the BER has dropped to zero, resulting in no further representation on the semiology scale.

Moreover, it is essential to analyze the BER response concerning the increase in the LLR clipping threshold, denoted as $r_{\max}$. Fig. 6 represents the plots between the BER and variation of clipping threshold $r_{\max}$, for different

E_b/N_0, under both lengths $N = 1024$ and $N = 512$, while maintaining rate $R = 1/2$, and quantization level is set to $q_s = 8$. The BER demonstrates a relatively direct proportionality to the variation of $r_{\max}$ up to a certain extent, beyond which it remains constant. It is important to note that excessively high clipping thresholds may also discard LLR values containing valuable information, potentially degrading the decoding accuracy. Therefore, a lower clipping threshold $r_{\max}$ is recommended to choose for effective decoding in both scenarios. For instance considering the code length $N = 512$ at $E_b/N_0 = 3.5$ the $r_{\max} = 4$ or 5 is considered a balanced tradeoff for efficient decoding accuracy and beyond this range, the BER starts to increase, indicating a degradation in decoding performance. It is crucial to find the appropriate balance between clipping threshold and decoding accuracy to ensure optimal performance in polar code decoding.

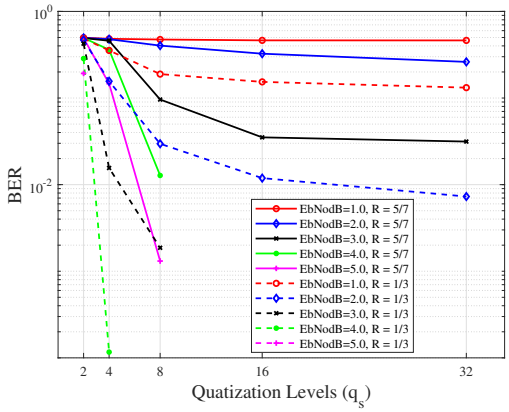


Fig. 4: *Representation of BER vs. quantization levels plots under different rates $R = 5/7$ and $R = 1/3$ in SC decoder of length $N = 1024$ and $r_{\max} = 3$.*

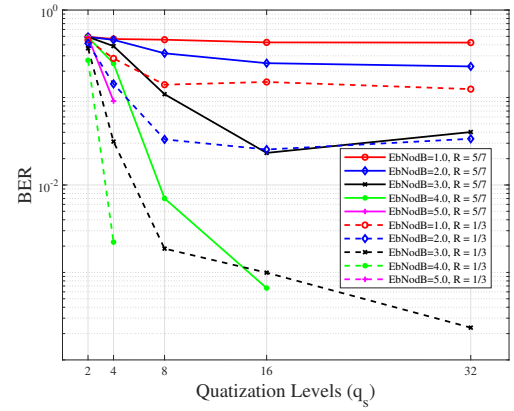


Fig. 5: *The BER vs. quantization levels plots under different rates $R = 5/7$ and $R = 1/3$ with SC decoder of length $N = 512$ and $r_{\max}$*

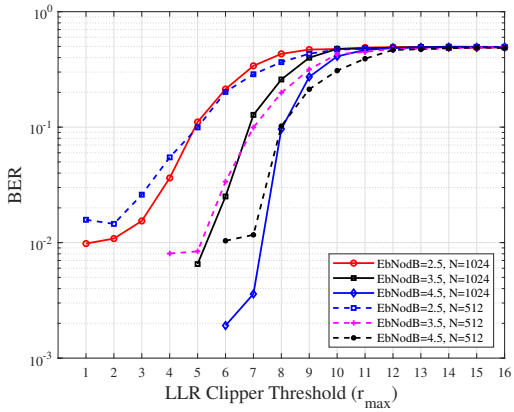


Fig. 6: *The plot between BER and variation of clipping threshold $r_{\max}$, at a constant code lengths $N = 1024$, $N = 512$, rate $R = 1/2$, and quantization level $q_s = 8$.*

5. CONCLUSION

This study explores the impact of LLR clipping and quantization on SC decoding performance in polar codes of different rates. By limiting received LLRs and quantizing them into different levels, we examine the tradeoff between decoding performance and quantization. We have observed, with the increase of quantization level the BER continues to fall down till $q_s = 16$ and afterward, it gets flattened, indicating negligible gain with increasing quantization levels. Similarly, a lower clipping threshold $r_{\max}$ is considered to effectively reduce impulsive noise and enhance decoding accuracy. The findings contribute to efficient decoding algorithms for practical applications, where clipping and quantization can be tailored to specific requirements.

LITERATURA

[1] B. Dai, R. Liu, and Z. Mei, "Novel non-linear demapper for soft decision decoder of ldpc codes in impulsive noise," *IET Communications*, vol. 13, no. 6, pp. 796–801, 2019.

[2] J. Wang, J. Li, S. Yan, W. Shi, X. Yang, Y. Guo, and T. A. Gulliver, "A novel underwater acoustic signal denoising algorithm for gaussian/non-gaussian impulsive noise," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 1, pp. 429–445, 2020.

[3] Y. Fu and H. Yang, "Improving log-likelihood ratio estimation with bi-gaussian approximation under multiuser interference scenarios," *Entropy*, vol. 23, no. 6, p. 784, 2021.

[4] M. Okumura, K. Asano, T. Abe, E. Okamoto, and T. Yamamoto, "Performance improvement of radiowave encrypted mimo communications using average llr clipping," *IEICE Transactions on Communications*, vol. 105, no. 8, pp. 931–943, 2022.

[5] H. B. Maad, A. Goupil, L. Clavier, and G. Gelle, "Clipping demapper for ldpc decoding in impulsive channel," *IEEE Communications Letters*, vol. 17, no. 5, pp. 968–971, 2013.

[6] K. Nikitopoulos and G. Ascheid, "Complexity adjusted soft-output sphere decoding by adaptive llr clipping," *IEEE Communications Letters*, vol. 15, no. 8, pp. 810–812, 2011.

[7] H. Rezaei, N. Rajatheva, and M. Latva-aho, "High-throughput rate-flexible combinational decoders for multi-kernel polar codes," *arXiv preprint arXiv:2301.10445*, 2023.

[8] D. B. Khalifa and M. Martel, "Constrained precision tuning," in *2022 8th International Conference on Control, Decision and Information Technologies (CoDIT)*, vol. 1. IEEE, 2022, pp. 230–236.

[9] C. Kim, D. Kam, S. Kim, G. Park, and Y. Lee, "Simplified ordered statistic decoding for short-length linear block codes," vol. 26, no. 8. IEEE, 2022, pp. 1720–1724.

[10] S. H. Hassani and R. Urbanke, "Polar codes: Robustness of the successive cancellation decoder with respect to quantization," in *2012 IEEE International Symposium on Information Theory Proceedings*. Ieee, 2012, pp. 1962–1966.

Systemy komórkowe 5G wspierane przez BSP, OZE oraz IPR

5G cellular systems supported by UAVs, RESs, and RISs

Adam Samorzewski[1]; Adrian Kliks[2,3];

[1] Politechnika Poznańska, Poznań, Polska, adam.samorzewski@doctorate.put.poznan.pl
[2] Politechnika Poznańska, Poznań, Polska, adrian.kliks@put.poznan.pl
[3] Uniwersytet Techniczny w Luleå, Luleå, Szwecja, adrian.kliks@ltu.se

DOI: 10.15199/59.2023.4.18

Streszczenie: W artykule rozważono zużycie energetyczne w sieciach komórkowych 5G zasilanych przez Odnawialne Źródła Energii (OZE) oraz wyposażonych w Inteligentne Powierzchnie Rekonfigurowalne (IPR) oraz Bezzałogowe Statki Powietrzne (BSP) *na uwięzi*, pełniące rolę mobilnych punktów dostępowych. Badania skoncentrowano na energetycznej stronie sieci dostępu radiowego (ang. *Radio Access Network* – RAN) zlokalizowanej na terenie miasta Poznania w Polsce. Zysk związany z wykorzystaniem generatorów OZE, czyli paneli fotowoltaicznych (ang. *photovoltaic panel* – PVP) dla stacji bazowych (ang. *base station* – BS) został przedstawiony w postaci dwóch współczynników: średniej liczby ładowań BSP (SLLB), aby zapewnić ciągły dostęp do usług mobilnych dla obsługiwanych terminali mobilnych użytkowników (ang. *user equipment* – UE) oraz średniej redukcji zużycia energii (SRZE) przez system bezprzewodowy.

Abstract: The paper considers energy consumption in 5G cellular networks powered by Renewable Energy Sources (RESs) and equipped with Reconfigurable Intelligent Surfaces (RISs) and tethered Unmanned Aerial Vehicles (UAVs), acting as mobile access points. The study was focused on the energy side of the Radio Access Network (RAN) located in the city of Poznań in Poland. The profit associated with the use of renewable energy generators, i.e. photovoltaic panels (PVPs) for base stations (BSs) is presented in the form of two factors: the average number of UAV charges (ANUC) to provide continuous access to mobile services for connected user equipment (UE) terminals, and the average reduction in energy consumption (AREC) of the wireless system.

Słowa kluczowe: 5G, Bezzałogowe Statki Powietrzne, Inteligentne Powierzchnie Rekonfigurowalne, Odnawialne Źródła Energii, sieci komórkowe

Keywords: 5G, cellular networks, Reconfigurable Intelligent Surfaces, Renewable Energy Sources, Unmanned Aerial Vehicles

1. WSTĘP

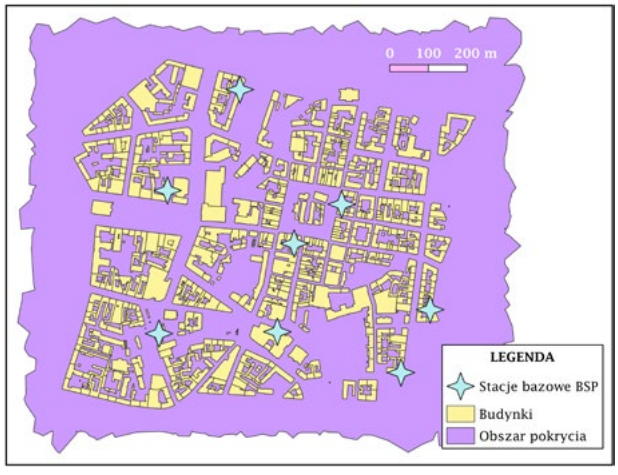
Obecne systemy telekomunikacyjne zasilane są głównie przez konwencjonalne źródła energii (KZE), tj., paliwami kopalnymi. Systemy te odpowiadają za około 25% ilości emisji dwutlenku węgla (CO_2) powodowanych przez segment ICT (ang. *Information and Communication Technology*), którego zapotrzebowanie na energię wydaje się być coraz większe z roku na rok [1]. Ponadto sektor ICT jest odpowiedzialny za sporą część globalnych emisji gazów cieplarnianych (ang. *Greenhouse Gas* – GHG), której możliwie niedoszacowana wartość może sięgać nawet od $2,1$ do $3,9\%$ [2]. Zatem, aby zredukować wspomnianą ilość gazów, konieczne może okazać się znalezienie alternatywnych źródeł energii, które zarówno zaspokoją zapotrzebowanie na energię systemów bezprzewodowych (4G, 5G oraz kolejnych generacji), jak i zmniejszą ilość CO_2 emitowanego do atmosfery. Wówczas zaangażowanie odnawialnych źródeł energii (OZE), np. paneli fotowoltaicznych, jest rozwiązaniem godnym uwagi ze względu na sposób pozyskiwania zasobów z promieniowania słonecznego, który sam w sobie wydaje się być mniej szkodliwy dla środowiska, tj., niezanieczyszczający. Co więcej, ilość potencjalnie generowanych zasobów energetycznych można by uznać w pewnym stopniu za niewyczerpalną (wyłączając konieczność wymiany paneli fotowoltaicznych po czasie ściśle określonym przez ich producenta) w kontekście długoterminowym [3, 4].

W ramach koncepcji systemów komórkowych 5-ej generacji usługi mobilne dzielą się na cztery główne grupy, z których trzy dotyczą odpowiednio wysokich przepływności, małych opóźnień oraz dużej liczby jednocześnie podłączonych urządzeń [5]. Z punktu widzenia realizacji pierwszej z nich, można napotkać ograniczenie związane ze skończoną liczbą bloków zasobów fizycznych (ang. *Physical Resource Blocks* – PRB) przypadającą na komórkę sieci. Problem ten można zauważyć zwłaszcza w obszarach miejskich, gdzie zaludnienie jest dość gęste. Zatem, aby zapewnić wystarczającą pojemność sieci bezprzewodowych na tego typu obszarach, zaproponowano rozwiązanie w postaci stacji bazowych obejmujących regiony tzw. małymi komórkami, które opiera się na rozmieszczaniu węzłów dostępowych małej mocy bardzo blisko siebie. Jednak przy wdrażaniu tej koncepcji ze względu na utrudnienia, takie jak niekorzystna architektura miejska czy braki finansowe na budowę nowych stacjonarnych stacji bazowych, istnieje ryzyko pojawienia się *luki sygnałowej*. Jednym ze sposobów uniknięcia tego zjawiska, coraz częściej branym pod uwagę w pracach naukowych, jest zaangażowanie dodatkowego sprzętu, np. bezzałogowych statków powietrznych (BSP) jako mobilnych stacji bazowych (ang. *Mobile Base Station* – MBS). To z kolei daje operatorom sieci komórkowych (ang. *Mobile Network Operator* – MNO) możliwość dynamicznego dostosowywania aktualnej lokalizacji węzłów dostępowych w celu pokrycia terenów, do których sygnał radiowy nie dociera, i/lub obsługi istniejącej infrastruktury telekomunikacyjnej w obszarach miejskich, gdzie liczba jednocześnie podłączonych

terminali użytkowników może ulegać wahaniom, a nawet przekraczać pierwotnie założoną pojemność systemu (np. z powodu wydarzeń publicznych) [6, 7].

W dodatku wykorzystanie tak zwanych inteligentnych powierzchni rekonfigurowalnych (IPR) do sterowania pokryciem radiowym systemów bezprzewodowych cieszy się dużym zainteresowaniem w aktualnej literaturze. IPR to urządzenie w postaci powierzchni składającej się z dużej liczby pasywnych (lub czasami aktywnych) elementów odbijających, które są w stanie samodzielnie wywołać pożądaną zmianę fazy i/lub amplitudy padającego na nie sygnału radiowego. Dzięki zapewnieniu elastycznej rekonfiguracji propagacji sygnału (po włączeniu IPR), operatorzy mobilni byliby w stanie osiągnąć lepszą wydajność swoich sieci poprzez zmniejszenie zakłóceń i przerw oraz podniesienie niezawodności, pojemności i przepływności łączy radiowych [8, 9].

W niniejszym artykule zbadano wydajność urządzeń BSP *na uwięzi* (przyłączonych przewodem do zewnętrznego zestawu bateryjnego umiejscowionego na ziemi/budynku) pracujących jako stacje bazowe, które są wyposażone zarówno w IPR (do potencjalnego wykorzystania w przyszłości), jak i OZE (do wydłużenia czasu pracy). Celem tej analizy było znalezienie najodpowiedniejszego rozmieszczenia i konfiguracji mobilnych stacji bazowych na określonym obszarze przy uwzględnieniu mocy nadawania, energii zużywanej przez BSP oraz produkcji energii przez OZE. Kolejno oszacowano kompromis między dodatkową masą potrzebną do przenoszenia elementów OZE i IPR a korzyściami związanymi z energią wytwarzaną przez OZE. Na koniec przeprowadzono symulacje opierając się na rzeczywistych danych atmosferycznych (zawarte w [10]), aby zweryfikować faktyczną wydajność proponowanego rozwiązania.



Rysunek 1: Mapa badanego obszaru (Poznań).

2. SCENARIUSZ SYSTEMOWY

Rozważany w pracy scenariusz uwzględnia sieć komórkową 5G zlokalizowaną w obrębie miasta Poznania w Polsce (dane z [11]), której stacjami bazowymi są bezzałogowe statki powietrzne zlokalizowane na podstawie rzeczywistych danych jednego z polskich operatorów sieci komórkowych (zawarte w [12]) i unoszące się 50 m nad ziemią. W przeprowadzonych badaniach założono, że rozmieszczenie dronów jest stacjonarne w czasie obsługi wspomnianego obszaru. Co więcej, na terenie sieci bezprzewodowej jest 100 losowo rozmieszczonych użytkowników, gdzie każdy z nich wymaga stałej przepływności równej 100 Mb/s na łączu w dół oraz 25 Mb/s na łączu w górę. Na Rys. 1 zamieszczono mapę przedstawiającą opisany powyżej scenariusz wdrożenia systemu bezprzewodowego.

2.1 Projektowanie sieci komórkowej

Rozważana sieć komórkowa została zaprojektowana przy użyciu specjalistycznego oprogramowania opisanego w [13] o nazwie *Green Radio Access Network Design* (GRAND). Wspomniane narzędzie optymalizuje również działanie sieci dostępu radiowego pod kątem zużycia energii i/lub narażenia ludzi na działanie pola elektromagnetycznego (PEM). Proces optymalizacyjny opiera się na wyznaczeniu optymalnej liczby włączonych komórek stacji bazowych (o predefiniowanych lokalizacjach) ze wszystkich dostępnych oraz dostosowanie ich parametrów transmisji sygnału radiowego w oparciu o chwilowe przepływności wymagane przez użytkowników podłączonych do systemu bezprzewodowego. Jako dane wejściowe oprogramowanie GRAND otrzymuje listy dostępnych stacji bazowych i aktywnych użytkowników oraz pliki 3D opisujące rozważany obszar pokrycia oraz znajdujące się na jego terenie budowle. Lokalizacja każdego użytkownika jest wybierana losowo przez symulator w każdym pojedynczym przebiegu. Dodatkowo na samym początku wszystkie układy nadawczo-odbiorcze (NO) węzłów dostępowych nadają sygnał radiowy z maksymalną mocą. Po rozmieszczeniu użytkowników na badanym terenie program wybiera, które komórki sieci mają zostać włączone i z jaką mocą powinny przesyłać dane, aby dotrzeć do jak największej liczby aktywnych użytkowników a jednocześnie by zminimalizować ich narażenie na promieniowanie PEM i/lub zużycie energetyczne systemu. Na tym etapie narzędzie GRAND ocenia potencjalne połączenia radiowe między siecią a każdym UE w oparciu zarówno o ich wymagania transmisyjne, jak i również maksymalnie dopuszczalną wartość tłumienia propagowanego sygnału przy założeniu różnych konfiguracji mocy nadawania. Następnie, w celu znalezienia optymalnego rozwiązania, formułuje się funkcję celu opierając się na przewidywanej optymalizacji narażenia ludzi na działanie PEM, zużycia energii lub obu. Następnie wszyscy użytkownicy rozpoczynają wymianę danych z odpowiednimi węzłami dostępowymi w sposób ciągły, powodując stałe obciążenie ruchem przez cały czas trwania pojedynczej symulacji. Równolegle wykonywane są obliczenia bilansu energetycznego dla wszystkich stacji bazowych, które następnie przekładają się na średnią liczbę potrzebnych doładowań akumulatorów na BSP z i bez OZE.

2.2 Wyposażenie

Implementacja paneli fotowoltaicznych w ramach oprogramowania GRAND została zainspirowana specyfikacją rzeczywistego urządzenia, którą można znaleźć w [14]. Przyjęto, że wykorzystane panele fotowoltaiczne są montowane na górnej części pokrywy BSP w postaci cienkowarstwowych ogniw słonecznych. W związku z tym, ze względu na znikomy wpływ na całkowity pobór mocy, ciężar paneli fotowoltaicznych został pominięty we wszystkich obliczeniach poświęconych charakterystykom energetycznym stacji BSP. Dodatkowo obliczenia te przeprowadzono dla czterech różnych dni (każdy rozpoczynający inną porę roku), aby pokazać, jak pozyskiwanie energii

przez panele fotowoltaiczne zależy od pory roku dla klimatu panującego Polsce.

Podobnie jak w przypadku paneli fotowoltaicznych, systemy bateryjne dla stacji mobilnych BSP również zostały zaprojektowane na podstawie rzeczywistej implementacji opisanej w [15]. Celem wyposażania dronów w akumulatory jest zaspokojenie ich zapotrzebowanie na energię, a także magazynowanie zasobów wytwarzanych przez generatory OZE. Rozładowanie ogniw baterii pociąga za sobą konieczność wymiany jej na nową, w pełni naładowaną. Niemniej jednak przyjęto założenie, że akumulatory można całkowicie rozładować, a 5% całkowitej energii zgromadzonej w jednym akumulatorze jest zawsze zużywane na loty urządzenia typu dron z i do stacji dokującej.

Układy nadawczo-odbiorcze węzłów dostępowych BSP pracują zgodnie z technologią MIMO (ang. *Multiple-Input-Multiple-Output*) wykorzystując 64 aktywne elementy antenowe (ang. *active antenna element* – AAE) transmitujące sygnały radiowe w paśmie częstotliwości 3500 MHz. Przyjęte w oprogramowaniu metody estymacji kanału i przetwarzania sygnału są zgodne ze schematem minimalnego błędu średnio-kwadratowego (ang. *minimum mean-squared error* – MMSE). Ponadto każda mobilna stacja bazowa posiada jedno urządzenie typu IPR z 16 identycznymi pasywnymi elementami odbijającymi, które realizują przesunięcie fazowe sygnału radiowego z rozdzielczością 6-bitową. W niniejszej pracy pominięto wpływ działania matryc IPR na propagację sygnału radiowego na badanym obszarze, w przeciwieństwie do ich oddziaływania na bilans energetyczny systemu bezprzewodowego. Jednakże wpływ urządzeń typu IPR na wydajność sieci komórkowej zostanie wzięty pod uwagę w przyszłych rozważaniach.

2.3 Modele energetyczne

Wszystkie modele energetyczne użyte do przeprowadzenia operacji matematycznych w GRAND (związanych zarówno z produkcją, jak i konsumpcją energii – prosumpcją) zostały bezpośrednio zaczerpnięte lub zainspirowane literaturą naukową lub rzeczywistymi wdrożeniami. Lista tych formuł została zawarta w Tab. 1.

Tabela 1: Lista modeli energetycznych

Nazwa modelu energetycznego	Oznaczenie	Źródło
Zużycie energii przez wielowirnikowy BSP	P_{BSP}	[16]
Zużycie energii przez MIMO układ NO	P_{MIMO}	[17]
Zużycie energii przez matrycę IPR	P_{IPR}	[8]
Moc wyjściowa panelu fotowoltaicznego	P_{PVP}	[18]
Składownie energii w systemie bateryjnym	P_{BAT}	[19]

3. KONFIGURACJA SYMULACJI

Kod źródłowy opracowanego oprogramowania został przygotowany w języku Java. Badanie opisanego powyżej scenariusza systemowego zostało przeprowadzone w formie 10 niezależnych przebiegów symulacyjnych, z których każdy uwzględnia 4 dni poprzedniego roku rozpoczynające różne jego pory – równonoc wiosenną $\left(20\text{ marca }2022\right)$, przesilenie letnie $\left(21\text{ czerwca }2022\right)$, równonoc jesienną $\left(23\text{ września }2022\right)$ oraz przesilenie zimowe $\left(21\text{ grudnia }2022\right)$. Parametry użytkowników (koordynaty współrzędnych oraz żądane warunki transmisyjne) są definiowane zawsze na początku każdego przebiegu symulacyjnego. Przyjęty krok czasowy odpowiada 1 minucie czasu symulacyjnego ($4\text{ dni} \cdot 24\text{ godziny} \cdot 60\text{ minut} = 5760$ kroków przypadających na pojedynczą symulację), w którym aktualizowane są dane dotyczące warunków pogodowych, a następnie wykonywane są obliczenia dotyczące generacji i zużycia energetycznego przez komponenty systemu bezprzewodowego.

Tabela 2: Bilans energetyczny dla BSP z i bez OZE

	Całkowita (i szczytowa) energia z OZE [Wh]	
	Bez OZE	*PVP*
Równonoc wiosenna	0 (0)	475,17 (60,57)
Przesilenie letnie	0 (0)	572,64 (91,86)
Równonoc jesienna	0 (0)	349,56 (65,15)
Przesilenie zimowe	0 (0)	17,67 (4,18)
Wartość średnia	0	353,76

	Średnia redukcja zużycia energii (SRZE) [%]	
	Bez OZE	*PVP*
Równonoc wiosenna	0	7,28
Przesilenie letnie	0	8,71
Równonoc jesienna	0	5,33
Przesilenie zimowe	0	0,27
Wartość średnia	0	5,4

	Średnia liczba ładowań BSP (SLLB)	
	Bez OZE	*PVP*
Równonoc wiosenna	9,03	8,16
Przesilenie letnie	9,06	8,1
Równonoc jesienna	9,03	8,4
Przesilenie zimowe	9,03	8,93
Wartość średnia	9,03	8,4

4. WYNIKI

Wyniki przeprowadzonych symulacji zostały załączone w Tab. 2. Pierwsza tabela przedstawia ilość energii, którą panele fotowoltaiczne mogą zebrać w ciągu całego roku dla pojedynczej stacji BSP, z wyszczególnieniem każdego sezonu. Wartości całkowite oraz szczytowe zawarte w tabeli wyznaczono w odniesieniu odpowiednio do pojedynczej doby i godziny. Zgodnie z początkowymi przewidywaniami, największą ilość zasobów, jaką mobilna stacja bazowa jest w stanie uzyskać z promieniowania słonecznego, przypada na przesilenie letnie (572,64 Wh), kiedy to szczytowa wartość procesu produkcji energii jest również najwyższa (91,86 Wh). Kolejno w rankingu uplasowały się wiosenna i jesienna równonoc oraz na końcu przesilenie zimowe. Można zatem zauważyć, że pod względem redukcji energii dostarczanej ze źródeł konwencjonalnych (tj. z akumulatorów ładowanych przez dedykowane stacje dokujące) przedstawionej w środkowej tabeli, kolejność jest adekwatna do wspomnianych powyżej zależności. Jednakże ze względu na ograniczenia związane z liczbą ogniw fotowoltaicznych oraz ich wydajnością wytwarzania energii, maksymalny uzyskany zysk energetyczny wyniósł 8,71% (przesilenie letnie). Należy również zauważyć, że dla kontrastu podczas przesilenia zimowego ten zysk jest znikomy (0,27%). Wreszcie ostatnia tablica prezentuje średnią liczbę ładowań stacji BSP po rozładowaniu jej systemu bateryjnego w ciągu doby. Ze względu na panujące warunki pogodowe, zróżnicowanie tej liczby można zaobserwować nawet wtedy, gdy mobilny punkt dostępowy nie jest zasilany przez generatory OZE. Najwyższą liczbę ładowań dronów odnotowano podczas przesilenia letniego (9,06), przy czym dla pozostałych dat wartość ta była jednakowa (9,03). Z drugiej strony, gdy urządzenie BSP jest obsługiwane przez panele fotowoltaiczne, podczas sezonu letniego wymaga ono najmniejszej średniej liczby ładowań bate-

rii (8, 1). Dość podobny zysk odnotowano dla równonocy wiosennej (8, 16), a kolejno za nią znalazła się równonoc jesienna, dla której liczba ładowań baterii dronu (z użyciem paneli solarnych) była równa jej wartości średniej dla pojedynczego dnia w roku (8, 4). Tym samym, ze względu na niemal zerowy wpływ wykorzystania paneli fotowoltaicznych w okresie przesilenia zimowego na charakterystykę bilansu energetycznego sieci, liczba ładowań stacji BSP wykorzystującej OZE zmniejszyła się w tym czasie w najmniejszym stopniu (8, 93) względem przypadku z ich pominięciem (9, 03).

5. WNIOSKI

Zaprezentowana w artykule praca podkreśla zalety związane z wykorzystaniem paneli fotowoltaicznych jako generatorów energii w sieciach komórkowych wyposażonych w urządzenia BSP *na uwięzi* oraz matryce IPR. Dla rozważanego scenariusza systemowego, ze względu na zarówno panujące w Polsce warunki pogodowe, jak i przyjęte konfiguracje stacji BSP oraz generatorów OZE, redukcja zużycia energii z KZE (i wynikające z tego oszczędności finansowe) wynoszą średnio 5, 4% rocznie w porównaniu do przypadku, w którym stacje bazowe systemu bezprzewodowego nie są zasilane z OZE. Chociaż odnawialne źródła energii, takie jak panele fotowoltaiczne, charakteryzują się zmiennymi w czasie i zależnymi od klimatu procesami pozyskiwania energii, to poprzez odpowiednie zarządzanie dostępnymi zasobami (radiowymi i energetycznymi) przy pomocy algorytmów zarządzania siecią (jak np. sterowanie ruchem telekomunikacyjnym, alokacja zasobów itp.) oraz dzięki dodatkowemu wyposażeniu, takiemu jak macierze IPR, istnieje możliwość poprawy uzyskanych rezultatów, a nawet zapewnienia autonomii energetycznej sieciom komórkowym bez pogarszania jakości dostarczanych usług mobilnych. Jednakże zarówno implementacja wspomnianych algorytmów, jak i badania ukierunkowane na ewaluację wpływu urządzeń typu IPR na propagację sygnału radiowego będą brane pod uwagę w przyszłych rozważaniach.

PODZIĘKOWANIA

Autorzy pragną podziękować prof. Margot Deruyck z Uniwersytetu w Gandawie — IMEC w Belgii za wsparcie niniejszej pracy poprzez udostępnienie programu GRAND. Praca została zrealizowana w ramach projektu nr 2021/43/B/ST7/01365 ufundowanego przez Narodowe Centrum Nauki (NCN) w Polsce.

LITERATURA

[1] C.-L. I et al, „Energy-efficient 5G for a greener future", Nature Electronics, vol. 3, pp. 182–184, 2020.

[2] C. Freitag et al., „The Climate impact of ICT: A review of estimates, trends and regulations", arXiv: 2102.02622v1.

[3] M. Deruyck et al, „Designing a multiple renewable energy source system to feed the wireless access network", Elsevier Sustainable Energy, Grids and Networks, vol. 31, 2022.

[4] M. K. Abid, M. V. Kumar, V. A. Raj, and M. D. K. Dhas, „Environmental Impacts of the Solar Photovoltaic Systems in the Context of Globalization", Ecological Engineering & Environmental Technology, vol. 24, no. 2, pp. 231–240, 2023.

[5] 3GPP, „Technical Specification Group Services and System Aspects; Release 15 Description; Summary of Rel-15 Work Items (Release 15)", TR 21.915 v15.0.0, 2019.

[6] M. Mozaffari, W. Saad, M. Bennis, Y. -H. Nam, and M. Debbah, „A Tutorial on UAVs for Wireless Networks: Applications, Challenges, and Open Problems", in IEEE Communications Surveys & Tutorials, vol. 21, no. 3, pp. 2334–2360, 2019.

[7] M. Alzenad, A. El-Keyi, F. Lagum, and H. Yanikomeroglu, „3-D Placement of an Unmanned Aerial Vehicle Base Station (UAV-BS) for Energy-Efficient Maximal Coverage", in IEEE Wireless Communications Letters, vol. 6, no. 4, pp. 434–437, 2017.

[8] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, „Reconfigurable Intelligent Surfaces for Energy Efficiency in Wireless Communication", in IEEE Transactions on Wireless Communications, vol. 18, no. 8, pp. 4157–4170, 2019.

[9] M. Di Renzo et al., "Smart Radio Environments Empowered by Reconfigurable Intelligent Surfaces: How It Works, State of Research, and The Road Ahead," in IEEE Journal on Selected Areas in Communications, vol. 38, no. 11, pp. 2450–2525, 2020.

[10] Visual Crossing, „Historical Weather Data & Weather Forecast Data", Źródło: https://www.visualcrossing.com/weather-data, Data: 21 maj 2023.

[11] System Informacji Przestrzennej (SIP), „Poznań – Model 3D", Źródło: http://sip.poznan.pl/model3d/#/legend, Data: 21 maj 2023.

[12] BTSearch, „Baza danych oraz mapa lokalizacji stacji BTS / pozwoleń UKE", Źródło: http://beta.btsearch.pl, Data: 21 maj 2023.

[13] G. Castellanos et al, „Multi-objective optimization of human exposure for various 5G network topologies in Switzerland", Computer Networks, vol. 2016, 2022.

[14] Solarland, „Solarland SLP020-12U – Specifications", Źródło: https://www.solar-electric.com/lib/windsun/SLP020-12U.pdf, Data: 21 maj 2023.

[15] Volt, „Volt Accumulator $LiFePO_4$ 12.8V 60Ah – Specifications", Źródło: https://voltpolska.pl/index.php?controller=attachment&id_attachment=320, Data: 21 maj 2023.

[16] S. Janji, A. Samorzewski, M. Wasilewska, and A. Kliks, „On the Placement and Sustainability of Drone FSO Backhaul Relays", in IEEE Wireless Communications Letters, vol. 11, no. 8, pp. 1723–1727, 2022.

[17] E. Björnson, J. Hoydis, and L. Sanguinetti, „Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency", *Foundations and Trends® in Signal Processing*, vol. 11, no. 3–4, pp. 154–655, 2017.

[18] HOMER Pro v3.15, „Documentation – HOMER's Calculations", Źródło: https://www.homerenergy.com/products/pro/docs/3.15/homers_calculations.html, Data: 21 maj 2023.

[19] Voltacon, „How long does it take to charge batteries from solar panels?", Źródło: https://voltaconsolar.com/blog/2021/04/27/how-long-does-it-take-to-charge-batteries-from-solar-panels, Data: 21 maj 2023.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

ANALIZA DYNAMICZNEGO WYBORU WEZŁA PRZETWARZAJĄCEGO Z ALOKACJĄ PRZEPŁYWÓW W PAKIETOWYCH SIECIACH XHAUL

ANALYSIS OF DYNAMIC PROCESSING NODE SELECTION WITH FLOWS ALLOCATION IN PACKET-SWITCHED XHAUL NETWORKS

Edyta Biernacka[1]; Mirosław Klinkowski[2];

[1] AGH Akademia Górniczo-Hutnicza Wydział Informatyki, Elektroniki i Telekomunikacji,
al. A. Mickiewicza 30, 30-059 Kraków edyta.biernacka@agh.edu.pl
[2] Instytut Łączności - Państwowy Instytut Badawczy,
ul. Swojczycka 38, 51-501 Wrocław, M.Klinkowski@il-pib.pl

DOI: 10.15199/59.2023.4.19

Streszczenie: Artykuł prezentuje wyniki wstępnej analizy dynamicznej alokacji zasobów dla rozproszonych jednostek przetwarzających (DU) wraz z alokacją wrażliwych na opóźnienia przepływów ruchu (fronthaul/midhaul) w konwergentnych sieciach xHaul z komutacją pakietów. Badane są trzy strategie alokacji, które różnią się kryterium wg którego dokonywany jest wybór węzła przetwarzającego (PP). Eksperymenty symulacyjne przeprowadzono dla sieci typu mesh z uwzględnieniem różnych limitów opóźnień przepływów oraz pojemności węzłów PP.

Abstract: The paper presents preliminary results of analysis of dynamic resource allocation for processing of distributed units (DU) with allocation of latency-aware traffic flows (fronthaul/midhaul) in convergent packet-switched xHaul networks. Three strategies are studied, which apply differ criteria for the selection of processing pool (PP) nodes. Simulation experiments have been conducted in a mesh network assuming differen flows latency limits and PP node capacities.

Słowa kluczowe: 5G, radiowe sieci dostępowe (RAN), pakietowe sieci xHaul, alokacja zasobów

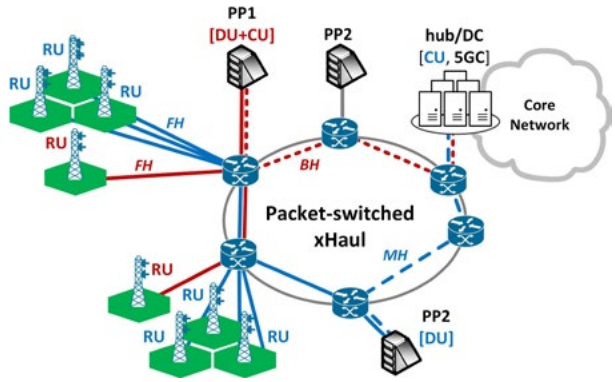
Keywords: 5G, radio access networks (RAN), packet-switched xHaul, resource allocation

1. WSTĘP

Sieci telekomunikacyjne podlegają gruntownym przeobrażeniom, które w dużej mierze wynikają z wdrażania sieci bezprzewodowych piątej generacji (5G) [1]. 5G umożliwia wszechobecny dostęp do nowych usług bezprzewodowych, klasyfikowanych jako: (a) rozszerzony mobilny dostęp szerokopasmowy (eMBB) dający wzrost dostępnej przepływności transmisyjnej urządzeń mobilnych (do 1-20 Gb/s), (b) masową komunikację maszyn (mMTC) umożliwiającą komunikowanie się urządzeń oraz realizację IoT (przewiduje się w sumie ponad bilion urządzeń, w tym ponad milion na kilometr kwadratowy) oraz (c) ultra-niezawodną komunikację o niewielkim opóźnieniu (URLLC), która jest niezbędna np. w aplikacjach wirtualnej i rozszerzonej rzeczywistości czy też w pojazdach autonomicznych.

Aby ułatwić wdrażanie usług 5G, specyfikacje techniczne architektury radiowej sieci dostępowej (5G RAN) [1] zdefiniowane przez organizację 3GPP umożliwiają podział funkcji przetwarzania podstawowego pasma radiowego między oddzielne jednostki, a mianowicie radiową (RU), rozproszoną (DU) i centralną (CU). Jednostki DU i CU mogą być zwirtualizowane i wykonywane na procesorach dostępnych w obiektach puli przetwarzania (PP) / centrach danych (DC) [10] zlokalizowanych w różnych miejscach sieci.

Jednym z głównych wyzwań przy wdrażaniu sieci 5G, w szczególności w przypadku "gęstych" sieci miejskich, jest budowa sieci xHaul [5, 11] realizujących transport ruchu pomiędzy elementami sieci RAN, łączącej radiowe punkty dostępowe (RU) z rozproszonymi elementami przetwarzającymi (PP) i rdzeniem sieci 5G (5G Core/5GC), zwykle zlokalizowanym w koncentratorze (hub). Sieci xHaul powinny wykorzystywać niedrogie technologie transmisyjne i jednocześnie być elastyczne tak by zapewnić łatwe przyłączenie i obsługę dużej liczby różnego typu radiowych punktów dostępowych. Technologie transportowe stosowane w sieciach 4G / LTE nie są w stanie sprostać powyższym wymaganiom. Stąd też jednym z proponowanych rozwiązań jest wykorzystanie dobrze znanej i taniej technologii Ethernet [3], w której dane przesyłane są w formie ramek / pakietów.



Rysunek 1: Pakietowa sieć xHaul.

Zastosowanie technologii Ethernet umożliwia implementację pakietowych sieci xHaul. W szczególności, stan-

dard IEEE 1914.1 [4] definiuje architekturę interfejsu fronthaul nowej generacji (NGFI), która zapewnia skalowalną i elastyczną łączność między elementami RAN za pomocą sieci xHaul z komutacją pakietów. NGFI umożliwia konwergentny (wspólny) transport różnych radiowych przepływów danych, takich jak fronthaul (RU-DU), midhaul (DU-CU) i backhaul (CU-5GC), między RU, PP i hubem/DC, w którym ruch jest agregowany, we współdzielonej fizycznej infrastrukturze sieciowej. Rysunek 1 przedstawia przykład sieci pakietowej xHaul.

Podstawowym wyzwaniem istniejącym w pakietowych sieciach xHaul jest zapewnienie odpowiedniego poziomu jakości usługi (QoS) dla realizowanych przepływów ruchu. Potrzeba takich gwarancji wynika zarówno z wysokich wymagań jakie stawiają usługi 5G jak i jest konieczna dla poprawnej realizacji funkcji przetwarzania radiowego w sieci RAN. Przepływy ruchu xHaul charakteryzują się zróżnicowanymi i ostrymi (w szczególności w przypadku ruchu FH) wymaganiami dotyczącymi dopuszczalnych opóźnień (czyli czasu potrzebnego na dostarczanie danych). Niedeterministyczny charakter transmisji w pakietowych sieciach xHaul powoduje, że pakiety przekierowywane jednocześnie do tego samego łącza w przełączniku wymagają buforowania. Buforowanie pakietów wprowadza opóźnienia, wpływając tym samym na jakość usługi. Stąd też przepływy ruchu w pakietowych sieciach xHaul wymagają odpowiedniej obsługi uwzględniającej ich wymagania QoS. Powinno się to odbywać poprzez alokację zasobów transmisyjnych i właściwe trasowanie ruchu w sieci (tzw. routing), mające na celu ominięcia przeciążonych, wprowadzających zbyt duże opóźnienia łączy, jak i odpowiednią alokację zasobów obliczeniowych w wybranych węzłach sieci, celem ich wykorzystania przez funkcje radiowe oraz dla realizacji usług 5G.

W niniejszym artykule dokonano analizy dynamicznego wyboru i alokacji węzłów PP wraz z alokacją przepływów w sieciach pakietowych xHaul. W badanym scenariuszu sieci, jednostki dostępu radiowego (RU) są wdrażane w danym obszarze sieci nie jednocześnie, ale sekwencyjnie, jak dla przykładu w przypadku rozbudowy sieci. Stąd, rozważany problem alokacji zasobów rozwiązywany jest dla poszczególnych, pojawiających się w czasie żądaniach przyłączenia RU, a nie dla całego zbioru RU jak np. w przypadku problemu planowania sieci [6, 7, 8].

W badaniach wykorzystano trzy strategie alokacji przepływów. Alokacja przepływów odbywała się na podstawie długości ścieżki do węzła przetwarzającego PP, obciążenia węzła PP oraz długości ścieżki do aktywnego węzła PP. W celu przeprowadzenia badań zaimplementowano symulator pakietowej sieci xHaul w środowisku C++/OMNeT++. Wyniki analiz w rozważanej sieci typu mesh wskazują, iż strategie oparte na priorytetowym wyborze węzłów aktywnych pozwalają na redukcję liczby aktywnych węzłów PP w sieci, co może prowadzić do zmniejszenie kosztu działania sieci, w porównaniu do strategii wyboru najbliższego PP.

2. SCENARIUSZ SIECI XHAUL

W tej pracy zakładamy podobny model sieci i ruchu jak w [8, 9]. Sieć implementuje architekturę NGFI [4], w której RU, PP i koncentrator (hub) są połączone za pomocą sieci xHaul z komutacją pakietów, wykorzystującej przełączniki Ethernet [3] do multipleksowania i routingu przepływów ruchu. Uwzględniane są obydwa kierunki transmisji, tzn. uplink (RU-PP-hub) i downlink (hub-PP-RU). Zakładamy alokację jednostek DU w węzłach PP i lokację jednostek CU w koncentratorze, co prowadzi do realizacji dwóch przepływów ruchu, mianowicie fronthaul (RU-PP) i midhaul (PP-hub), dla każdego kierunku transmisji. Stąd też przyłączenie RU do sieci wiąże się z koniecznością alokacji czterech przepływów ruchu. Rozważane w analizie w Sekcji 4 maksymalne dopuszczalne opóźnienia dla przepływów FH to 100 μs oraz 50 μs [4], podczas gdy limit opóźnień dla przepływów MH wynosi 1 ms.

W [12], dane radiowe są enkapsulowane [2] i wysyłane okresowo (co 66, 6 μs) jako ramki Ethernet. Każda ramka ma stały rozmiar równy 1542 bajtów. Dane w postaci paczek są przesyłane ze stałymi przepływnościami przy założeniu pełnego wykorzystania zasobów radiowych. Dane te są buforowane w przełącznikach i następnie przesyłane w kolejności swoich priorytetów, tzn., strumienie FH mają wyższy priorytet niż strumienie MH.

W szacowaniu opóźnień przepływów uwzględniamy zarówno opóźnienia statyczne, jak i dynamiczne, którym ulegają ramki Ethernet podczas transmisji w sieci [6, 7]. Opóźnienia statyczne uwzględniają opóźnienia propagacji w łączach sieciowych (zakładając prędkość 2×10^5 km/s), opóźnienia w przechowywaniu i przesyłaniu dalej w przełącznikach (5 μs na przełącznik) oraz czasy transmisji impulsów w łączach (w zależności od długości paczki danych i przepływności łącza). Dynamiczne opóźnienia reprezentują opóźnienia buforowania paczek danych w łączach wyjściowych przełącznika i do ich oszacowania stosujemy model obliczeń opóźnień "najgorszego przypadku" opisanego w [6, 8, 9].

3. DYNAMICZNY WYBÓR WĘZŁA PP Z ALOKACJĄ PRZEPŁYWÓW

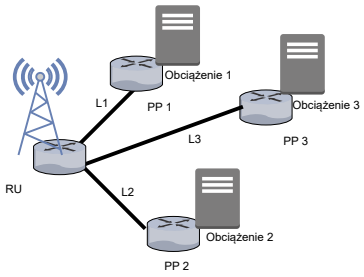
Analizowany w pracy problem dynamicznej alokacji zasobów w pakietowej sieci xHaul dotyczy ustanawiania połączeń pomiędzy jednostkami końcowymi (tzn. RU, PP oraz hub) dla napływających sekwencyjnie zgłoszeń dołączenia węzłów RU do sieci (np. w kontekście jej rozbudowy). Problem ten sprowadza się do zagadnienia wyboru węzła PP, w tym alokacji zasobów obliczeniowych dla przetwarzania DU, oraz alokacji przepływów FH/MH w sieci pomiędzy wybranymi lokalizacjami w obydwu kierunkach transmisji (uplink/downlink). Alokacja jest możliwa jeżeli dostępne są zasoby obliczeniowe do obsługi RU (tzn. alokacji DU) w wybranym węźle PP, dostępne są zasoby transmisyjne we wszystkich łączach należących do ścieżek routingu realizujących przepływy FH/MH, oraz opóźnienia przepływów FH/MH, zarówno alokowanych jaki i już obecnych w sieci, nie przekraczają dopuszczalnych limitów.

W pracy analizujemy trzy strategie wyboru węzła PP z alokacją przepływów:

- *NearestPP* - dla danego RU wybierany jest najbliższy możliwy do alokacji węzeł PP, tzn. o najkrótszej ścieżce routingu pomiędzy RU-PP i taki, że możliwa jest alokacja zasobów PP oraz zasobów transmisyjnych w łączach na ścieżkach routingu dla przepływów FH i MH (w obydwu kierunkach transmisji) i opóźnienia wszystkich przepływów nie przekraczają dopuszczalnych limitów;
- *MaxUsedPP* - wybierany jest najbardziej obciążony węzeł PP dla którego możliwa jest alokacja zasobów przetwarzania i transmisyjnych z zapewnieniem QoS przepływów; gdy alokacja nie jest możliwa, wybierany jest najbliższy nieaktywny węzeł PP;

- *NearestActivePP* - wybierany jest najbliższy aktywny węzeł PP dla którego możliwa jest alokacja zasobów przetwarzania i transmisyjnych z zapewnieniem QoS przepływów; gdy alokacja nie jest możliwa, wybierany jest najbliższy nieaktywny węzeł PP;

Rysunek 2 ilustruje działanie powyższych strategii na przykładzie pojedynczego RU z trzema węzłami PP.



Rysunek 2: Przykład pojedynczego RU z trzema dostępnymi węzłami PP.

Zakładając długość ścieżek $L1 < L2 < L3$, nieaktywny węzeł PP1 (brak obsługiwanych RU) oraz obciążenie PP2 < PP3, metoda *NearestPP* dokona wyboru węzła PP1 z alokacją zasobów dla przepływów FH wzdłuż ścieżki RU-PP1, metoda *MaxUsedPP* dokona wyboru węzła PP3 z alokacją zasobów wzdłuż ścieżki RU-PP3, natomiast *NearestActivePP* dokona alokacji zasobów wzdłuż ścieżki RU-PP2, przy spełnieniu QoS.

4. ANALIZA SYMULACYJNA

W tej części przedstawiono założenia eksperymentów symulacyjnych oraz wyniki wraz z analizą. Do przeprowadzenia badań został przygotowany symulator zestawiania połączeń w środowisku OMNeT++ dla sieci pakietowej xHaul. Kody programu zostały napisane w języku C++. W pierwszym etapie zaimplementowano funkcje wczytania topologii sieci (parametry sieci pakietowej xHaul) i generowania żądań (wg. modelu klient-sieć) dla dynamicznego scenariusza sieci. Dla dynamicznego zestawiania połączeń, pojedyncze żądanie zostało podzielone na przepływy (fronthaul/midhaul) w zależności od typu węzłów końcowych (RU/PP/hub) pomiędzy którymi odbywa się transmisja danych radiowych. Zostały zaimplementowane funkcje dynamicznie sprawdzające całkowite obciążenie węzłów obliczeniowych PP, wielkość natężenie ruchu w łączach sieci, całkowite opóźnienie przepływów oraz algorytm wyboru najkrótszej ścieżki routingu. Zostały zestawione połączenia dla przepływów fronthaul/midhaul w sieci pakietowej xHaul.

W przeprowadzonych symulacjach użyto sieci WRO-17 [9], przedstawionej na Rysunku 3. Topologia WRO-17 została opracowana dla celów badawczych na podstawie podzbioru rzeczywistych lokalizacji anten radiowych (łącznie 79 RU) w centrum Wrocławia w Polsce, gdzie przełączniki są umieszczone w pobliżu anten i połączone za pomocą łączy poprowadzonych wzdłuż ulic. Długości wszystkich łączy w WRO-17 odzwierciedlają rzeczywiste fizyczne długości przedstawionych połączeń i zostały zwiększone o pewną losową wartość, od 100 do 200 metrów, aby uwzględnić doprowadzenie światłowodu do urządzeń PP i przełączników.

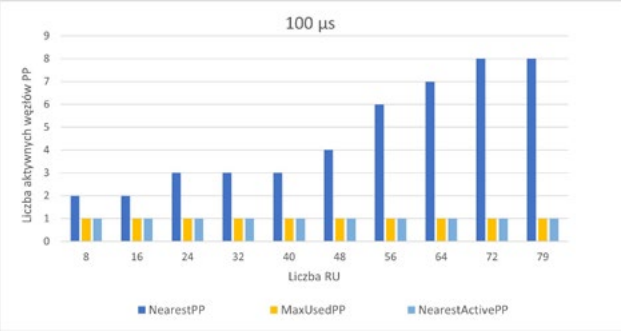
W sieci założyliśmy następującą przepustowość łączy 50 Gbit/s dla RU, 200 Gbit/s dla łączy pomiędzy przełącznikami, 800 Gbit/s dla węzłów PP i koncentratora (hub). Jest to dwukrotnie większa przepustowość niż w [9], jednakże przy takich założeniach w sieci nie pojawiły się blokady żądań. Celem przedstawionych badań jest sprawdzenie liczby aktywnych węzłów PP w sieci dla proponowanych metod dynamicznej alokacji przepływów. Przyjęte w analizie przepływności xHaul, rozważane w [8, 9], przedstawiono w Tabeli 1. Podobnie jak w [7, 8], alokacja przepływów odbywała się na najkrótszych ścieżkach.

Rysunek 3: Topologia sieci WRO-17; RU, PP, przełączniki i koncentrator (hub) są oznaczone odpowiednio trójkątami, kwadratami, okręgami i sześciokątem.

W pierwszym eksperymencie symulacyjnym założyliśmy nieograniczoną pojemność węzłów PP. Rysunki 4 oraz 5 przedstawiają wyniki liczby aktywnych węzłów PP w funkcji liczby RU w sieci WRO-17 dla limitu opóźnień przepływów FH, odpowiednio $L^{\mathrm{FH}} = 100$ μs i $L^{\mathrm{FH}} = 50$ μs.

Tabela 1: Analizowane przepływności xHaul [Gbit/s].

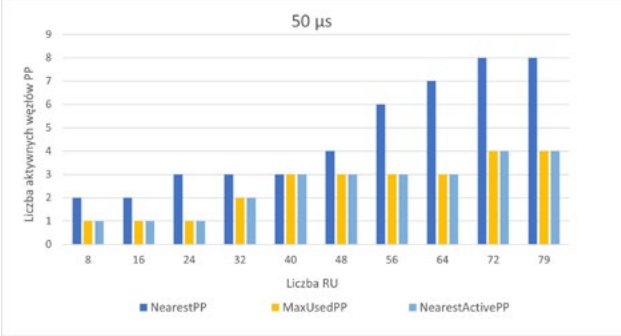
	Fronthaul	Midhaul
Uplink	5.496	0.774
Downlink	6.076	1.016



Rysunek 4: Liczba aktywnych węzłów PP w funkcji liczby przyłączonych RU przy założeniu nieograniczonej pojemności węzłów PP dla $L^{\mathrm{FH}} = 100$ μs.

Można zauważyć, że metoda *NearestPP* dokonuje aktywacji większej liczy PP w porównaniu do *MaxUsedPP* i *NearestActivePP*. Dla liczby RU równej 72 metoda *NearestPP* aktywuje wszystkie węzły PP. Wyjaśnić to można faktem wyboru najkrótszej ścieżki do węzła PP przez

metodę *NearestPP*, bez uwzględnienia aktywności i obciążenia PP. *MaxUsedPP* i *NearestActivePP* dążą w pierwszej kolejności do wyboru aktywnych węzłów PP, dlatego liczba aktywnych węzłów PP jest mniejsza dla tych metod. Większy budżet opóźnień ($L^{\mathrm{FH}} = 100$ μs) pozwala na aktywację mniejszej liczby PP niż w przypadku $L^{\mathrm{FH}} = 50$ μs, dzięki możliwości wyboru dalszych PP.



Rysunek 5: Liczba aktywnych węzłów PP w funkcji liczby przyłączonych RU przy założeniu nieograniczonej pojemności węzłów PP dla $L^{\mathrm{FH}} = 50$ μs.

Następnie zbadaliśmy wpływ dostępnej pojemności węzłów PP, wyrażonej jako liczba możliwych do obsłużenia RU w danym PP, na działanie analizowanych metod. Symulacje przeprowadziliśmy dla $L^{\mathrm{FH}} = 100$ μs i przy założeniu przyłączonych 79 RU. Uzyskane wyniki, zaprezentowane na Rysunku 6, potwierdzają poprzednie wnioski i pokazują wpływ dostępnej pojemności PP na liczbę aktywnych PP w sieci.

Rysunek 6: Liczba aktywnych węzłów PP w funkcji pojemności PP (wyrażonej jako maksymalna liczba obsługiwanych RU) zakładając 79 przyłączonych RU dla $L^{\mathrm{FH}} = 100$ μs.

5. WNIOSKI

Rozpatrywany w pracy problem dotyczył dynamicznej alokacji zasobów przetwarzających (DU) wraz z alokacją wrażliwych na opóźnienia przepływów ruchu (FH/MH) w sieciach xHaul, w których transmisja danych opiera się na komutacji pakietów. Badany przypadek dotyczył scenariusza sieci 5G RAN, w którym jednostki dostępu radiowego (RU) są wdrażane w danym obszarze sieci nie jednocześnie, ale sekwencyjnie (np. w kontekście rozbudowy sieci). Analizowano wpływ różnych strategii wyboru węzła PP na ilość aktywnych w sieci węzłów PP uwzględniając różne limity opóźnień dla ruchu FH i różne pojemności węzłów. Przedstawione wstępne wyniki badań wskazują na skuteczność strategii polegającej na priorytetowym wyborze jednego z już aktywnych wezłów PP. Jednocześnie nie zaobserwowano różnic pomiędzy strategią wyboru najbliższego węzła aktywnego i węzła najbardziej obciążonego.

Należy zauważyć, że w rozważanej sieci nie zaobserwowano blokady połączeń, co wynikało głównie z dużych pojemności łączy pozwalających na alokację wszystkich przepływów. Zastosowanie łączy o mniejszych pojemnościach wpłynie zarówno na ograniczenie dostępności zasobów transmisyjnych jak i zwiększenie opóźnień pakietów ze względu na większe czasy transmisji, w tym dłuższe oczekiwanie w buforach przełączników. Stąd, w kolejnym kroku opracowane będą efektywne metody wyboru DU i alokacji przepływów dla takich właśnie przypadków. Także, w dalszych pracach planowana jest analiza dynamicznych sieci xHaul, w których natężenie przepływów ruchu ulega zmianie w zależności od potrzeb przy uwzględnieniu dostępnych zasobów oraz zapewnieniu gwarancji opóźnień.

PODZIĘKOWANIA

Praca powstała w wyniku realizacji projektu badawczego OPUS o nr 2018/31/B/ST7/03456 finansowanego ze środków Narodowego Centrum Nauki.

LITERATURA

[1] 3GPP. 2022. „Architecture description (release 17)” Tech. Spec. TS 38.401, v17.0.0, Sophia Antipolis, Francja.

[2] eCPRI. 2019. „Common public radio interface: eCPRI V2.0 interface specification”.

[3] IEEE. 2018. „802.1cm-2018 – IEEE standard for local and metropolitan area networks – time-sensitive networking for fronthaul”.

[4] IEEE. 2020. „IEEE 1914.1-2019 - IEEE standard for packet-based fronthaul transport networks”.

[5] C.-L. I i inni. 2018. „RAN revolution with NGFI (xhaul) for 5G”. IEEE J. Lightw. Technol., 36 (2) : pp. 541–550.

[6] Klinkowski Mirosław. 2020. „Optimization of latency-aware flow allocation in NGFI networks”. Computer Communications, 161 : 344–359.

[7] Klinkowski Mirosław. 2020. „Latency-aware DU/CU placement in convergent packet-based 5G fronthaul transport networks”. Applied Sciences, 10 (21).

[8] Klinkowski Mirosław. 2023. „Analysis of Latency-Aware Network Slicing in 5G Packet xHaul Networks”. Intl Journal of Electronics and Telecommunications, 69 (2) : 335–340.

[9] Mroziński Damian, Klinkowski Mirosław, Walkowiak Krzysztof. 2023. „Cost-Aware DU Placement and Flow Routing in 5G Packet xHaul Networks”. IEEE Access, 11 : 12710-12726.

[10] Xiao, Y., Zhang J., Ji Y. 2020. „Can fine-grained functional split benefit to the converged optical-wireless access networks in 5G and beyond?”. IEEE Trans. Netw. Serv. Manag., 17 (3) : 1774–1787.

[11] O-RAN Alliance. 2020. „Xhaul packet switched architectures and solutions”. Tech. Spec. v1.0.

[12] Perez G. O., Larrabeiti D., Hernandez J. A. 2019. „5G new radio fronthaul network design for eCPRI-IEEE 802.1CM and extreme latency percentiles”. IEEE Access, vol. 7, pp. 82218–82229.

BADANIE TERMINALA 5G DO ZASTOSOWAŃ IoT

TESTING 5G TERMINAL FOR IoT APPLICATIONS

Tomasz Kosiło[1]; Jarosław Marski[2]; Karol Radecki[3]

[1] Instytut Radioelektroniki i Technik Multimedialnych, Politechnika Warszawska, Warszawa, tomasz.kosilo@pw.edu.pl
[2] T – Mobile Polska S.A., Warszawa, jaroslaw.marski@t-mobile.pl
[3] Instytut Radioelektroniki i Technik Multimedialnych, Politechnika Warszawska, Warszawa, karol.radecki@pw.edu.pl

DOI: 10.15199/59.2023.4.20

Streszczenie: **W pracy przedstawiono wyniki badań terminala 5G do zastosowań IoT. Badano parametry nadawanego sygnału, jego moc, poziomy emisji poza pasmowych, parametry modulacji oraz właściwości odbiornika. Pomiary wykonano przy pomocy testera radiokomunikacyjnego R&S CMX500. Wyniki odniesiono do obowiązujących norm ETSI, 3GPP. We wnioskach przedstawiono wpływ parametrów odbieranego sygnału na właściwości użytkowe terminala.**

Abstract: **The paper presents the results of research on the 5G terminal for IoT applications. The parameters of the transmitted signal, its power, out-of-band emission levels, modulation parameters and receiver properties were examined. Measurements were made using the R&S CMX500 radio communication tester. The results were compared to the applicable ETSI and 3GPP standards. The conclusions show the influence of the parameters of the received signal on the usable properties of the terminal.**

Słowa kluczowe: **Sieci komórkowe, 5G StandAlone, Non StandAlone, terminale IoT, pomiary terminala.**

Keywords: **Mobile networks, 5G StandAlone, Non StandAlone, IoT terminals, terminal measurements.**

1. WSTĘP

Zastosowania Internetu Rzeczy (IoT) czy też transmisji M2M szybko rosną. Znaczenie ma tu rozwój transmisji danych w sieciach komórkowych, a szczególnie możliwości sieci 5G. W 2016 r. 3GPP opublikowała pierwsze standardy transmisji do zastosowań IoT (E-UTRA, wersja 13) w sieciach komórkowych. W tym okresie powstały m. in. NB-IoT i LTE-M dostosowane do sieci 4G, które są kompatybilne z 5G i w najbliższej przyszłości będą stosowane w sieciach 5G do obsługi masowych sieci czujników generujących niewiele danych, tanich, o małym poborze mocy zasilania.

5G przewiduje trzy kategorie zastosowań:

- mMTC (massive Machine-Type Communication) dla masowych urządzeń, nadających małe ilości danych w trudnych warunkach, o małym poborze mocy
- URLLC (Ultra-Reliable, Low Latency Communications) niezawodna transmisja dla przemysłowych i krytycznych aplikacji IoT
- eMBB (enhanced Mobile Broadband) dla masowych zastosowań konsumenckich

NB-IoT i LTE-M odpowiadają kategorii mMTC. Natomiast kategoria URLLC otwiera nowe możliwości tworzenia zastosowań przemysłowych, systemów transportu, samochodowych i wielu innych [8]. Różne opracowania np. [1], [2] wskazują na znaczy wzrost inwestycji i zastosowań systemów IoT w najbliższym czasie. Do 2025 r. systemy IoT mają generować ponad połowę światowej ilości danych.

Obecnie nie ma standardów IoT stworzonych dla sieci 5G. Dostępne są jednak terminale transmisji danych 5G, które mogą być podstawą nowych rozwiązań. IoT jest często deklarowane jako główne przeznaczenie takich terminali. W pracy przebadano jeden z takich terminali.

System 5G może pracować w pasmach FR1 (450 MHz do 6 GHz) i FR2 (24,25 GHz do 52,6 GHz); ze względu na typ sterowania łączem radiowym może to być tryb SA (StandAlone) i NSA (Non StandAlone). NSA wykorzystuje sieć LTE do sygnalizacji (anchor), a kanał 5G do transmisji danych. Terminale w trybie NSA pracują w zdefiniowanej kombinacji pasm EN-DC – czyli kombinacji pasm LTE i 5G. Przykładem popularnej kombinacji EN-DC wykorzystywanej obecnie przez operatorów w Polsce to B3-N1 (1800MHz LTE + 2100MHz 5G). W trybie SA nie jest wykorzystywany kanał LTE, transmisja i zarządzanie łączem odbywa się w kanale 5G. W 5G występują dwa rodzaje separacji dupleksowej: FDD oraz TDD, w zależności od wykorzystanego kanału. Tabela kanałów i częstotliwości z pasma FR1 z podziałem na typ dupleksu jest zamieszczona w [4].

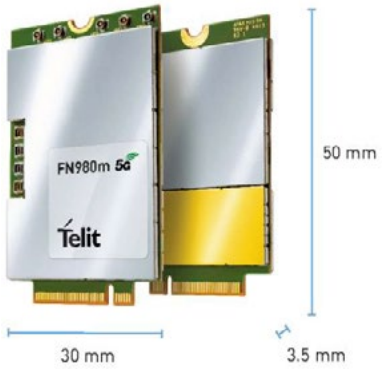
2. UŻYTE URZĄDZENIA

2.1. Badany terminal

Badano terminal FN980m firmy Telit, którego głównym przeznaczeniem wg producenta są systemy IoT. Terminal ma format karty M2. Może pracować w sieciach 3G, 4G i 5G. W sieciach 5G obsługuje pasma FR1 (sub-6 GHz) n1, n2, n3, n5, n7, n12, n14, n20, n28, n30, n41, n66, n71, n77, n78, n79 i FR2 (mmWave) n257, n258, n260, n261. W sieci 4G obsługuje pasma LTE: 1, 2, 3, 4, 5, 7, 8, 12, 13, 14, 17, 18, 19, 20, 25, 26, 28, 29DL, 30, 32, 34, 38, 39, 40, 41, 42, 43, 46(LAA), 48(CRS), 66, 71.

W sieci 5G terminal zapewnia transmisję do: 5,5 Gb/s DL / 2,7 Gb/s UL, a w sieci LTE do: 2 Gb/s DL / 211 Mb/s UL.

Moc nominalna nadajnika wynosi 23 dBm. W sieciach 5G oraz 4G terminal może pracować w trybach FDD oraz TDD (na wybranych kanałach) - natomiast w sieci 3G pracuje w trybie FDD. Typowa szerokość pasma nadajnika w trybie FDD wynosi 20MHz, a w trybie TDD od 20MHz do 100MHz. Czułość odbiornika jest w zakresie od -97dBm do -110dBm.



Rys. 1. Terminal FN980m. [7]

Terminal FN980m jest zaprojektowany zgodnie ze standardem 3GPP Rel. 15, który umożliwia pracę w trybach SA oraz NSA. W terminalu zastosowano procesor Qualcomm® Snapdragon™ X55, zoptymalizowany dla zastosowań IoT/eMBB.

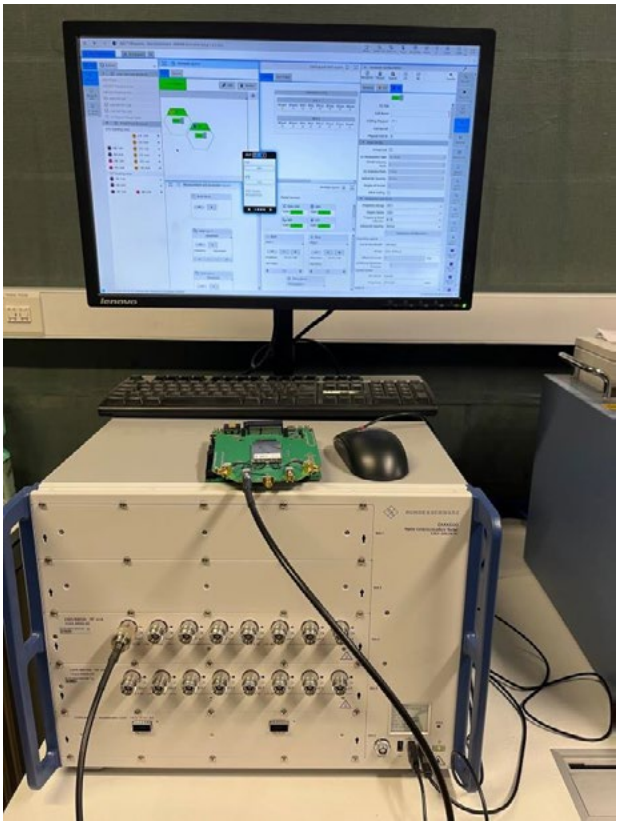
2.2. Układ pomiarowy

Pomiary terminala wykonano za pomocą testera radiokomunikacyjnego – emulatora sieci 5G typu CMX500 firmy Rohde & Schwarz [6].

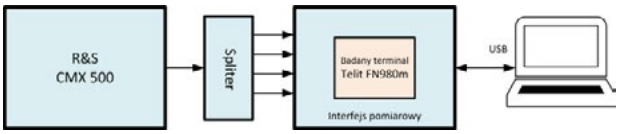
Tester pracuje w zakresie częstotliwości LTE oraz 5G: FR1 i w rozbudowanej wersji w zakresie milimetrowym FR2. Obsługuje pracę 5G w trybach SA i NSA. Umożliwia pomiary parametrów nadajnika i odbiornika badanego terminala. Wyniki pomiarów są zobrazowane na ekranie podłączonego monitora.

Rys. 2. Tester Rohde & Schwarz CMX 500 [6].

Wygląd stanowiska pomiarowego pokazano na Rys. 3. Terminal jest połączony z testerem poprzez płytę TLB (zielona płytka), która służy do celów pomiarowo testowych. Gniazda na płycie pozwalają na dołączenie kabli; użyto kabli współosiowych 50Ω o długości 50cm. Płyta zwiera też interfejs USB. Terminal sterowano z komputera PC przez interfejs USB, za pomocą komend AT, które wysyłano z aplikacji PuTTY.



Rys. 3. Wygląd stanowiska pomiarowego.



Rys. 4. Schemat układu pomiarowego.

2.3. Wyniki pomiarów

Do pomiarów wykorzystano tester 5G R&S CMX500 w układzie pomiarowym jak na Rys. 4. Przyjęto następujące parametry emulowanej sieci:

- Tryb NSA (Non StandAlone)
- Kombinacja EN-DC: LTE (B3) 5G (B78)

Oznacza to parametry sieci jak podano w Tab. 1.

Tester emuluje środowisko sieciowe i usługowe e2e. Poza elementami sieci szkieletowej zawiera serwer usługowy o adresie IPv4 172.100.20.201, na którym są możliwe do uruchomienia usługi takie jak FTP czy iPERF.

Terminal działał w trybie automatycznym, automatycznie dostosowywał się do parametrów emulowanej sieci takich jak pasmo częstotliwości, technologia dostępowa (LTE, 5G) oraz tryb pracy SA lub NSA

W trakcie pomiarów realizowano pogorszenie jakości połączenia w dół dla kanału 5G, poprzez obniżanie całkowitej mocy wyjściowej nadajnika symulowanej (przez CMX500) stacji bazowej. W trakcie badań obserwowano następujące parametry: blokową stopę błędów (BLER), przepływność 5G dla łącza w dół w Mb/s jak i w procentach odniesionych do maksymalnej szybkości uruchomionego strumienia danych. Z terminala odczytano poziom sygnału odbieranego RSRP, jakość łącza w dół

RSRQ, średni czas transmisji pakietów Ping o rozmiarze 32 bajtów.

Uruchomiono testowy strumień danych o przepływności 23Mb/s, który był przesyłany zarówno przez sieć LTE jak i 5G w proporcjach ca 3Mb/s LTE i 19Mb/s w 5G. Sprawdzano również czas przejścia pakietów Ping.

Pomiar Ping wykonano na komputerze PC wykorzystywanym do sterowania badanym terminalem. Uruchomiono IP thethering w oparciu o interfejs rNDIS (czyli modemowy dostęp do sieci z użyciem protokołu IPv4). Wysyłano pakiety Ping na adres testowanego serwera, działającego w symulatorze, za pomocą linii komend uruchomionej na komputerze PC. Zapisywano średnią wartość opóźnienia dla 4 pakietów o rozmiarze 32 bajtów.

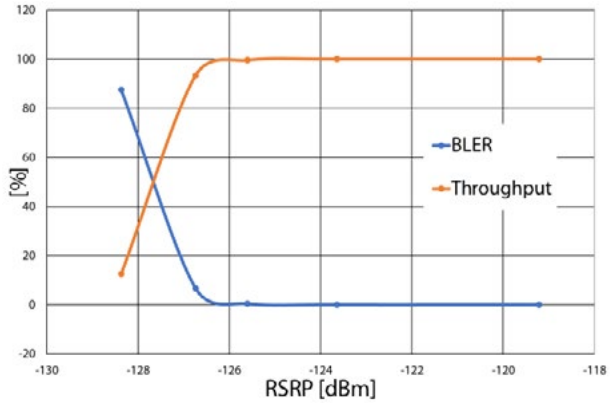
Tab. 1. Parametry badanego łącza 5G

Parametry LTE		
Pasmo	B3	
Dupleks	FDD	
Liczba Resource Blocks	50	
Szerokość pasma	10 MHz	
	Downlink	Uplink
f środkowa [MHz]	1842,5	1747,5
Kanał EARFCN	1575	19575
MIMO	SISO 1xN	SISO 1xN
Moc [dBm]	-57,2	
Modulacja		16 QAM
Parametry 5G		
Pasmo	N78	
Częstotliwość	3549,990 MHz	
Kanał NR AFRCN	636666	
Dupleks	TDD	
odstęp podnośnych OFDM	30 kHz	
Szerokość kanału	100 MHz	
Modulacja	64 QAM	
MIMO	SISO 1xN	
Moc w zakresie	-90 -99 dBm	

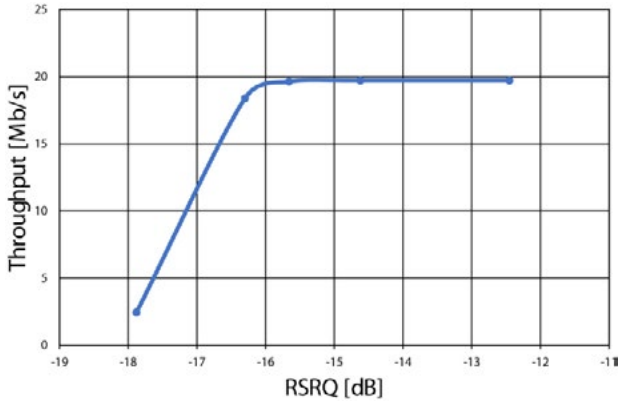
Z wyników pomiarów pokazanych na Rys. 5 i Rys. 6 wyraźnie wynika wpływ jakości połączenia na przepływność bitową transmisji oraz na średni czas przesyłania pakietu Ping.

Wyniki pokazują, że przy poziomie RSRP 125,6 dBm pojawia się blokowa stopa błędów BLER na poziomie 0,4%, której zależność w całym badanym zakresie jest odwrotnie proporcjonalna do wyrażonej w procentach uzyskanej przepływności bitowej w kanale 5G.

Wyniki BLER dla kolejnych punktów pomiarowych RSRP pokazują znaczący wzrost BLER w funkcji obniżającego się poziomu sygnału. Dla poziomu -128,37 dBm uzyskano BLER 87,51% i jednocześnie określono ten punkt jako granicę „zasięgu". Przy sygnale niższym o 0,5dB połączenie utraciło stabilność.

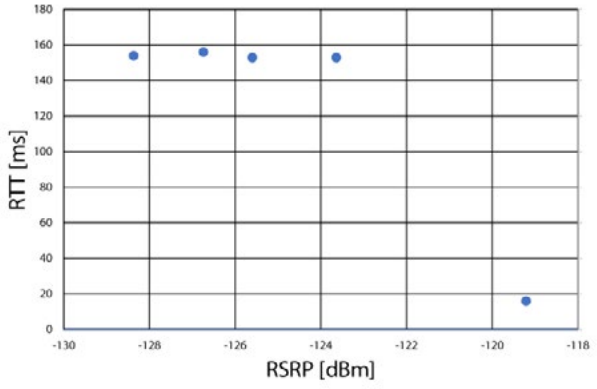


Rys. 5. Wynik pomiaru blokowej stopy błędów i przepływności w zależności od mocy sygnału (RSRP - Reference Signal Received Power [5]).



Rys. 6. Przepływność w zależności od jakości połączenia (RSRQ - Reference Signal Received Quality [5]).

Jedną z obserwacji jest wysoka stabilność szybkości bitowej dla parametrów odbioru w zakresie RSRP powyżej poziomu -125dBm, jednakowoż dla poziomów poniżej -119dBm nastąpiło znaczące zwiększenie opóźnienia pakietów Ping, bez obserwacji dalszego pogorszenia, przy kolejnych mniejszych poziomach sygnału odbieranego, co by mogło wskazywać na to, że pakiety Ping były przesyłane w większości przez kanał LTE.



Rys. 7. Średnie opóźnienie pakietów Ping. Każdy pomiar zawierał 4 pakiety.

Mierzono również niektóre parametry nadawanego sygnału w paśmie FR1:

1. przebieg widma mocy w paśmie nadajnika oraz poziomy mocy w kanałach, bezpośrednio sąsiadujących,

2. poziomy emisji poza pasmowych,
3. EVM, błędy fazy i częstotliwości oraz IQ offset dla kilku wartościowości modulacji QAM

Uzyskane wyniki mieściły się w granicach normy 3GPP [3].

Sieci 5G mogą znajdować zastosowanie w aplikacjach IoT, wymagających dużych szybkości przesyłania danych np. do transmisji sygnałów wizyjnych. Nasze badania wskazują, że przebadane łącze terminala 5G może być wykorzystywane do przesyłania sygnałów wizyjnych, ale dla zapewnienia stabilnego połączenia należałoby po stronie aplikacji zastosować adaptacyjny kodek wideo, aby można było dostosować przekazywany obraz do zmieniających się parametrów transmisji danych w wyniku zwiększania odległości terminala od stacji bazowej. Ponadto tryb NSA daje dodatkowe zabezpieczenie w postaci ciągłego połączenia z siecią LTE, co w przypadku mniejszego zasięgu dla częstotliwości 3,5 GHz, zapewni stabilne połączenie dla mniejszych szybkości transmisyjnych, gdzie sygnał 5G będzie zbyt słaby lub niedostępny.

Znaczącym odbiorcą usług IoT/5G będzie szeroko rozumiany transport (smart mobility) obejmujący monitorowanie pojazdów, sterownie ruchem, obsługę parkingów, pojazdy autonomiczne [9]. Istotny obszar zastosowań IoT/5G to także inteligentne miasta (smart city), co zapewni zwiększenie funkcjonalności infrastruktury, poprawę jakości życia i bezpieczeństwa mieszkańców [10]. Kolejnym przykładem zastosowań IoT/5G jest medycyna. Możliwość zbierania i analizy ogromnej ilości danych udoskonali metody diagnostyczne. Szybkość i niezawodność transmisji 5G zapewni nawet zdalne wykonywanie złożonych operacji [11].

3. PODSUMOWANIE

W pracy przedstawiono wyniki badań niektórych parametrów nadajnika i odbiornika istotnych dla praktycznego zastosowania terminala w systemie IoT. Wszystkie zmierzone parametry nadajnika i odbiornika mieściły się w granicach określonych normą 3GPP. W pracy skupiono się na zasymulowaniu warunków brzegowych dla odbioru radiowego dla najwyższej badanej częstotliwości 5G, ponieważ przekroczenie granicy zasięgu na tych pasmach nastąpi w rzeczywistych warunkach propagacyjnych wcześniej niż dla niższych częstotliwości. Oszacowano graniczną wartość BLER dla sygnału 5G (87,51%) jako granicę zasięgu odbieranego sygnału, zbadano warunki występowania wysokiej stabilności szybkości bitowej oraz małego opóźnienia pakietów Ping w sieci. Zmierzone opóźnienia nie przekraczają 100ms, które dla zastosowań IoT w radiowych systemach monitoringu procesów przemysłowych oraz bezpieczeństwa drogowego są wystarczające. Potrzebne są dalsze badania terminala, zwłaszcza zasięgu sygnału w rzeczywistym terenie oraz opóźnienia krytycznego dla różnych zastosowań IoT.

LITERATURA

[1] 5G Americas. White paper. 2019.“5G – The Future of IoT”. https://www.5gamericas.org/5g-the-future-of-iot/

[2] GSMA White paper. 2020. “5G IoT Private & Dedicated Networks for Industry 4.0”. https://www.gsma.com/iot/resources/5g-private-npn-industry40/

[3] ETSI. 2023. “5G; NR; User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone”. TS 138 101-1 v17.8.0

[4] ETSI. 2021. “5G; NR; User Equipment (UE) conformance specification; Radio transmission and reception; Part 1: Range 1 standalone”. TS 138 521-1 V16.8.0

[5] ETSI. 2020. “5G; NR; Physical layer measurements”.TS 138 215 V16.2.0

[6] Rohde & Schwarz. 2022. “5G device application testing with the R&S®CMX500 radio communication tester”

[7] Telit. 2021. “FN980 Family Hardware Design Guide, Telit Technical Documentation VV0301603 Rev. 7”

[8] Schulz Philipp et al. 2017. “Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture”. IEEE Communications Magazine, February: 70-78.

[9] Karagiannis Georgios, Klein Thomas. 2020. “IoT Relation and Impact on 5G”. AIOTI WG03, Relase 3.0, April: 1-141.

[10] Eshrat Md., Alahi E., et al. 2023. “Integration of IoT-Enabled Technologies and Artificial Intelligence (AI) for Smart City Scenario: Recent Advancements and Future Trends”. Sensors (23) 5206: 1-36.

[11] Devi Delshi H, et al. 2023. “5G Technology in Healthcare and Wearable Devices: A Review”. Sensors (23) 2519: 1-21.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

UCZENIE MASZYNOWE W SIECIACH KWANTOWEJ DYSTRYBUCJI KLUCZA

MACHINE LEARNING IN QUANTUM KEY DISTRIBUTION NETWORKS

Wojciech Szczepanik; Marcin Niemiec

Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie,
Mickiewicza 30, 30-059 Kraków; wojciech.szczepanik@agh.edu.pl

DOI: 10.15199/59.2023.4.21

Streszczenie: We współczesnych sieciach komputerowych bezpieczeństwo przesyłania informacji jest jednym z priorytetów. Z tego względu kwantowa dystrybucja klucza – rozwiązanie odporne na ataki z użyciem komputera kwantowego – zyskuje coraz większą popularność. Usprawnienie działania sieci za to odpowiedzialnych jest obecnie tematem intensywnych badań. W artykule przedstawiono możliwe zastosowania uczenia maszynowego w sieciach kwantowych. Opisano możliwe scenariusze, wyzwania oraz możliwości jakie otwiera ten kierunek badań.

Abstract: In modern computer networks, the security of information transmission is one of the priorities. For this reason, quantum key distribution – a solution resistant to quantum computer attacks – is gaining popularity. Improving the performance of the networks responsible for this is currently a topic of intense research. The article presents possible applications of machine learning in quantum networks. Possible scenarios, challenges and opportunities opened up by this direction of research are described.

Słowa kluczowe: kwantowa dystrybucja klucza, uczenie maszynowe, kryptografia, cyberbezpieczeństwo

Keywords: quantum key distribution, machine learning, cryptography, cybersecurity

1. WSTĘP

Wymiana informacji w cyfrowej formie jest jednym z kluczowych aspektów funkcjonowania obecnego świata. Konieczność zapewnienia bezpieczeństwa komunikacji jest coraz większa w związku ze zwiększającą się liczbą ataków. Cyberbezpieczeństwo odgrywa w tym kluczową rolę. Pozwala ono na szacowanie i kontrolę ryzyka, wykorzystywanie algorytmów kryptograficznych w celu zachowania poufności i integralności danych, odpowiednie projektowanie infrastruktury oraz wiele innych działań.

Bezpieczeństwo obecnej komunikacji w dużym stopniu bazuje na kryptografii asymetrycznej w celu ustalania wspólnego klucza dla szyfru symetrycznego. Zakłada ona, że jedynie dwie strony komunikujące się ze sobą będą wiedziały jaki jest ostatecznie ustalony klucz kryptograficzny. Przykładem takiego algorytmu jest algorytm Rivesta-Shamira-Adlemana (RSA) [1]. Jego bezpieczeństwo opiera się na problemie faktoryzacji liczb złożonych. Jednak jednym z poważnych wyzwań współczesnej kryptografii są postępy w pracach nad opracowaniem komputerów kwantowych, które będą w stanie złamać algorytm RSA [2]. Problem dotyka również pozostałych obecnie stosowanych algorytmów kryptografii asymetrycznej.

Wprawdzie obecnie RSA jest wciąż uznawany za bezpieczny, ale nie można przewidzieć jak długo tak pozostanie. Dodatkową motywacją do przyspieszenia badań przeciwdziałających temu zagrożeniu jest metoda deszyfrowania retrospektywnego. Potocznie jest ona zwana również zbierz teraz, odszyfruj później. Zatem jeśli chcemy zapewnić poufność danych w długiej perspektywie (np. informacje medyczne pacjentów), już teraz powinniśmy stosować algorytmy kryptograficzne odporne na atak z udziałem komputera kwantowego. Z tej przyczyny prowadzone są wzmożone prace nad nowymi rozwiązaniami wraz z ich implementacją.

Potencjalne rozwiązania można podzielić na dwie grupy – wykorzystujące kryptografię postkwantową oraz kwantową dystrybucję klucza (QKD). Pierwsza z nich wykorzystuje algorytmy odporne na ataki z wykorzystaniem komputera kwantowego. Jednak jej problemem są rozmiary kluczy – znacznie większe niż obecnie stosowanych algorytmów. Druga z nich zapewnia bezpieczeństwo na poziomie mechaniki kwantowej. Dzięki temu komunikacja przy użyciu bitów kwantowych, na której opiera się QKD zapewnia bardzo wysoki poziom bezpieczeństwa, gdyż umożliwia wykrywanie nawet pasywnego podsłuchu. Posiada ona jednak pewne wady – np. wymaga zastosowania dedykowanego kanału do komunikacji kwantowej. Poza tym, obecnie działające rozwiązania QKD mają problemy z wydajnością oraz posiadają różne inne ograniczenia/wyzwania.

Usprawnień działania sieci kwantowej dystrybucji kluczy można dokonać na wiele sposobów. Wprawdzie zwiększenie liczby dostępnych łączy pozwala zwiększyć przepustowość takich sieci, ale wymaga znaczących nakładów finansowych. Zastosowanie tradycyjnych algorytmów optymalizacji może również usprawnić działanie sieci, ale należy pamiętać że są one jednak często kosztowne numerycznie. Takie podejście zwiększa ryzyko pojawienia się dodatkowych opóźnień w reakcji na zmiany. Wykorzystanie heurystycznych metod opartych na uczeniu maszynowym może przyczynić się do ograniczenia obecnych problemów z sieciami QKD, jednocześnie generując koszty i efekty uboczne mniejsze od innych rozwiązań [3].

Artykuł składa się z czterech rozdziałów. Rozdział 2 przedstawia obszary w których można zastosować uczenie maszynowe wraz z dotychczasowymi publikacjami. W

kolejnym – rozdziale 3 – pokazano przyczynę trudności w prowadzeniu badań wraz z propozycją ich rozwiązania przez zastosowanie symulatorów sieci kwantowych. W ostatnim rozdziale podsumowano dotychczasowe postępy wraz z wnioskami odnośnie przeprowadzonych badań i przyszłych zastosowań uczenia maszynowego w sieciach kwantowej dystrybucji klucza.

2. OBSZARY ZASTOSOWAŃ

Sieci kwantowej dystrybucji klucza (QKDN), zwane także sieciami kwantowymi, mogą zostać usprawnione przez rozwiązania oparte na sztucznej inteligencji. Usprawnienia te polegają na poprawie wydajności, która jest niezwykle istotna w przypadku sieci o dużych rozmiarach. W [3] zasugerowano podział rozwiązań na trzy obszary: dotyczące łącza kwantowego, systemów zarządzania kluczem oraz warstwie sterowania i zarządzania.

2.1 Łącze kwantowe

Pierwszym zastosowaniem w obszarze łącza kwantowego jest predykcja wydajności kanału. Z prędkością transmisji kluczy powiązane są takie parametry jak liczba pojedynczych fotonów wykrytych przez detektor czy kwantowa bitowa stopa błędów (QBER). Celem minimalizacji błędów spowodowanych przez niedoskonałości kanału można wykorzystać algorytmy sztucznej inteligencji. Pozwalają one przewidzieć jakich zmian należy dokonać w kanale, aby zmaksymalizować prędkość transmisji.

Przykładem takiego zastosowania jest tzw. bezzapachowy filtr Kalmana (UKF) zastosowany w [4]. Kwantowa dystrybucja klucza oparta na zmiennych ciągłych (CV-QKD) wykorzystuje modulację amplitudy oraz fazy. Wraz z sygnałem kwantowym przesyłany jest sygnał pilota, którego faza była wykrywana. Przeprowadzone testy z użyciem UKF na komercyjnym sprzęcie z wykorzystaniem 20-kilometrowego łącza stwierdziły możliwość uzyskania transmisji dla stosunku sygnału do szumu sygnału pilota wynoszącego nawet 4dB.

Próbę usprawnienia kwantowej dystrybucji klucza opartej na zmiennych ciągłych dokonano też w [5]. Wykorzystując sieć neuronową w połączeniu z optymalizacją Bayesowską uzyskano wyniki potwierdzające wysoką niezawodność rozwiązania z prawdopodobieństwem wynoszącym ponad 99%. W porównaniu z metodą numeryczną osiągnięto 10^7 większą szybkość działania.

Optymalizacja parametrów transmisji w systemie QKD jest kolejnym zagadnieniem do którego rozwiązania zasugerowano uczenie maszynowe. Parametrami określamy m. in. takie wielkości jak liczba fotonów użytych do przesłania danych w kanale kwantowym czy użyty protokół. Ze względu na ograniczone możliwości obliczeniowe systemu QKD działanie standardowych metod optymalizacji nie jest realizowane w czasie rzeczywistym. Uczenie maszynowe może rozwiązać ten problem.

Temat wyboru protokołu transmisji poruszono w [6]. Decyzja między BB84, TF-QKD i MDI-QKD może zostać podjęta przez klasyczne algorytmy takie jak wyszukiwanie lokalne. Ich działanie jest jednak zbyt wolne, by mogło działać w czasie rzeczywistym. Dlatego podjęto próbę zastosowania uczenia maszynowego. Porównano kilka algorytmów wśród których najlepszym okazał się być losowy las decyzyjny (RF). Osiągnął on dokładność na poziomie 98%.

W [7] zajęto się tym samym problemem. Tutaj wykorzystano jednak nietestowaną wcześniej architekturę sieci neuronowej – Tree-CNN. Porównano ją z innymi algorytmami i uzyskano AUC wynoszące blisko 100%. Wśród porównywanych algorytmów znalazł się RF z poprzednio wymienionego artykułu. Okazał się jednak mniej skuteczny od zaproponowanej Tree-CNN.

Predykcja żywotności komponentów systemu jest kolejnym obszarem badań, który może pomóc w usprawnieniu działania sieci QKD. Ze względu na zewnętrzne czynniki środowiskowe i czynniki wewnętrzne nie jest możliwe jednoznaczne stwierdzenie momentu, w którym komponenty systemu przestaną działać. Zbierane liczne dane z czujników (takie jak temperatura, moc, energia elektryczna itp.) są trudne do przeanalizowania i określenia przewidywanej żywotności. W takich przypadkach za rozwiązanie mogą posłużyć sieci neuronowe. Według wiedzy autorów nie zostały jeszcze opublikowane wyniki badań nad przewidywaniem żywotności komponentów sieci QKD z wykorzystaniem algorytmów uczenia maszynowego.

2.2 Warstwa zarządzania kluczem

Po przesłaniu kluczy kanałem kwantowym następuje konieczność ich przechowania i dystrybuowania do aplikacji/usług ich wymagających. Zastosowanie uczenia maszynowego w zarządzaniu kluczem ma na celu poprawę wydajności i stabilności systemów zarządzania kluczem. Dodatkowo umożliwia również detekcję anomalii i prób ataków.

Zanim klucze zostaną zmagazynowane agent zarządzenia kluczami musi wybrać odpowiednią długość. Przyczyną takiej konieczności jest zapotrzebowanie algorytmów na różne długości kluczy, np. AES-256 zawsze wymaga 256 bitów, natomiast szyfr z kluczem jednorazowym (OTP) wymaga liczby bitów równej długości przesyłanej wiadomości. Możliwe jest przekonwertowanie kluczy w razie potrzeby, lecz wiąże się to z powstaniem opóźnień. Dlatego użycie uczenia maszynowego może być pomocne do przewidywania najbliższych zapotrzebowań. Według wiedzy autorów nie zostały jeszcze opublikowane wyniki badań nad wyborem formatu przechowywania kluczy wykorzystujących algorytmy uczenia maszynowego. Jest to jednak bardzo obiecujący kierunek badań, w którym autorzy widzą duży potencjał, przez co rozpoczęli współpracę z badaczami z kilku europejskich ośrodków.

Zadanie zarządzania magazynem klucza polega na planowaniu i wydajnym dysponowaniu zasobami kluczy. Moduł zarządzania jest odpowiedzialny za odbieranie i przechowywanie kluczy wygenerowanych przez moduły QKD oraz dostarczanie tych kluczy do aplikacji kryptograficznych według ich zapotrzebowań. Z jednej strony, jeśli klucze będą zbyt długo przechowywane to powinny zostać zniszczone. Z kolei jeśli chwilowe zapotrzebowanie przez aplikacje będzie zbyt duże, bufor może zostać całkowicie opróżniony co uniemożliwi obsługę wszystkich żądań. Ze względu na poziom złożoności całego procesu sugerowane jest zastosowanie uczenia maszynowego do oceny zdrowia magazynu klucza i podejmowanie na tej podstawie decyzji odnośnie udzielania kluczy.

Porównanie sieci neuronowej z prostymi heurystykami zostało przeprowadzone w [8]. W stworzonym w tym celu symulatorze porównano wszystkie algorytmy dla dwóch topologii sieci. Wykorzystując uczenie ze wzmocnieniem (RL) przeprowadzono trening sieci neuronowej. Zaproponowany algorytm okazał się we wszystkich sce-

nariuszach lepszy niż pozostałe heurystyki zarówno pod względem mniejszego prawdopodobieństwa blokowania jak i większego wykorzystania dostępnych kluczy.

Detekcja potencjalnych ataków w warstwie zarządzania kluczami to kolejny interesujący kierunek badawczy. Wykrywanie podejrzanych zachowań w warstwie zarządzania kluczem dokonywane jest na podstawia analizy żądań przesyłanych przez aplikacje. Przesyłanie poprawnych żądań w nadmiarowej ilości przez przejętą przez atakującego aplikację może doprowadzić do wyczerpania kluczy dla innych, prawidłowo działających aplikacji (DoS). W celu wykrywania takich anomalii oraz blokowaniu takich żądań można wykorzystać uczenie maszynowe. Według wiedzy autorów dotychczas opublikowane badania nad detekcją anomalii w warstwie zarządzania kluczem nie wykorzystywały algorytmów uczenia maszynowego.

2.3 Warstwa sterowania i zarządzania

Warstwa zarządzania jest odpowiedzialna za zbieranie informacji o funkcjonowaniu całego systemu. Dane są pozyskiwane z wielu źródeł co sprawia, że ich formaty znacząco różnią się od siebie. Konieczne jest oczyszczenie danych, zmniejszenie ich nadmiarowości i doprowadzenie ich do użytecznej formy. Tradycyjne algorytmy nie są wystarczające do skutecznej klasyfikacji i agregacji danych. Wykorzystanie uczenia maszynowego może pozwolić na przetworzenie danych o znaczących rozmiarach. Tak przygotowane informacje mogą zostać wykorzystane przez inne algorytmy, w tym takie, które wykorzystują uczenie maszynowe. Według wiedzy autorów nie zostały jeszcze opublikowane wyniki badań nad wstępnym przetwarzaniem logów z wykorzystaniem sztucznej inteligencji.

Kiedy dwa niesąsiadujące ze sobą jednostki chcą rozpocząć między sobą komunikację muszą pozyskać od kontrolera sieci QKD informację o trasie przekazu klucza. Ten otrzymując w żądaniu informację o źródle i celu oraz liczbie wymaganych przez aplikacje kluczy odpowiada za jej wyznaczenie. Kontroler otrzymuje informację o stanie sieci z warstwy zarządzania. Wśród tych informacji można wymienić wskaźniki konsumpcji kluczy czy pozostałą liczbę kluczy w węzłach. Tradycyjne algorytmu routingu mogą wyznaczyć optymalne trasy jednak są złożone obliczeniowo. Ponieważ sytuacja w sieci zmienia się dynamicznie takie metody posiadałyby zbyt duże opóźnienia. W związku z tym sugerowane jest wybieranie tras przekazu klucza w sposób heurystyczny przez wykorzystanie algorytmów uczenia maszynowego.

Routing wykorzystujący uczenie maszynowe przetestowano w [9]. Rozwiązanie było świadome jaki jest stan buforów w sieci i na podstawie danych z pomiarów wyznaczało trasy pojawiających się zapotrzebowań. Osiągnięte rezultaty okazały się lepsze od dotychczasowo stosowanych heurystyk.

Dodatkowym zadaniem warstw sterowania i zarządzania jest reagowanie na awarie. Uszkodzenie łącza bądź węzła sieci wchodzącego w trasę przekazu klucza powoduje przerwanie transmisji. Taka sytuacja wpływa również na stan magazynu klucza w innych węzłach. Zdolność predykcji takich wydarzeń pozwoliłaby na przeciwdziałanie ich skutkom, które zagrażają bezpieczeństwu transmisji. Pozwoliłoby to np. na zmienienie obecnych tras na takie, które pominą elementy sieci mogące w najbliższym czasie ulec awarii. Sama predykcja może zostać dokonana przez algorytmy uczenia maszynowego zdolne do ekstrakcji cech. Według wiedzy autorów nie zostały jeszcze opublikowane wyniki badań nad predykcją i reagowaniem na awarie w warstwach sterowania i zarządzania wykorzystujące algorytmy uczenia maszynowego.

3. SYMULACJE

Opracowanie i analizowanie skuteczności algorytmów obsługujących sieci QKD wymaga dużej ilości danych. Bazowanie na większej ilości informacji zapewnia lepsze i pewniejsze wyniki. Jednak pozyskiwanie zbiorów danych jest kosztownym procesem, zwłaszcza jeśli te dane mają być zbliżone do rzeczywistych.

W mniej złożonych sytuacjach możliwe jest wykorzystanie fizycznych urządzeń. Uruchomienie kwantowego łącza punkt-punkt pozwala na wytrenowanie i przetestowanie algorytmów uczenia maszynowego skupiających swoje działanie wyłącznie na pojedynczym łączu. Ma to zastosowanie w przypadku redukcji bitowej stopy błędów czy wyborze optymalnego protokołu transmisji. Zbudowanie większych sieci lub przetestowanie rozwiązań jeszcze nie zaimplementowanych stanowi poważniejszy problem.

Rozwiązaniem które warto rozważyć jest wykorzystanie symulatorów. Uruchamiane na stacjach roboczych lub serwerach pozwalają na przybliżone do rzeczywistego działanie sieci bez konieczności posiadania jej fizycznych elementów. Proces symulacji jest jednak kosztowny obliczeniowo, a samo opracowanie symulatora jest niezwykle skomplikowane oraz czasochłonne. Większość obecnie dostępnych symulatorów skupia się na działaniu warstwy fizycznej w połączeniach punkt-punkt. Na szczególną uwagę zasługują dwa rozwiązania pozwalające na symulację rozległych sieci QKD i które dzięki swojej dokładności mają zastosowanie do celów badawczych.

Pierwszym z nich jest QKDNetSim [10]. Został zaimplementowany jako moduł w środowisku ns-3, które jest symulatorem sieci zdarzeń dyskretnych. QKDNetSim implementuje klucze QKD, bufory QKD, postprocesowe aplikacje i szyfry. System zarządzania kluczem wspiera rozwiązania ETSI GS QKD 014 i ETSI GS QKD 004. Moduł jest nadal rozwijany, najnowsza wersja kodu nie jest jednak upubliczniona. Obecnie publicznie dostępna jest starsza wersja oprogramowania [11]. Istnieje możliwość przetestowania oprogramowania przez stronę internetową [12].

Drugim symulatorem jest QuISP – Quantum Internet Simulation Package [13]. Został zaimplementowany jako pakiet w OMNeT++, który również jest symulatorem sieci zdarzeń dyskretnych. QuISP wspiera m.in. złożone topologie, skanowanie łączy podczas inicjalizacji symulacji oraz wszystkie podstawowe funkcjonalności OMNeT++. Oprogramowanie jest nadal rozwijane. Wraz z kodem źródłowym [14] zawarte są informacje o planach implementacji kolejnych funkcjonalności.

4. PODSUMOWANIE

W związku ze wzrostem mocy obliczeniowych dostępnych komputerów kwantowych konieczne jest opracowanie algorytmów kryptograficznych odpornych na ataki z ich udziałem. Wśród dwóch możliwych rozwiązań należy wymienić kryptografię postkwantową oraz sieci kwantowej dystrybucji klucza. Druga z tych metod zapewnia bezpieczeństwo oparte na zasadach mechaniki kwantowej. Nadal trwają badania nad tą technologią, a obecne implementacje posiadają ograniczoną funkcjonalność. Jednak usprawnienia działania sieci QKD można dokonać przez

algorytmy uczenia maszynowego. Pozwoli to na zwiększenie wydajności lub wyższy poziom bezpieczeństwa oraz przyspieszy praktyczną implementację takich rozwiązań.

Sieci kwantowej dystrybucji klucza posiadają wiele różnorodnych obszarów, w których można użyć uczenia maszynowego. Począwszy od poprawy wydajności w kanale fizycznym, przez zarządzanie dostępnymi kluczami, do monitorowania i sterowania całą siecią. Ponieważ problemy w każdym z tych obszarów się różnią, konieczne jest przygotowanie wielu algorytmów rozwiązujących każdy z nich. Obecnie prowadzone badania z wykorzystaniem uczenia maszynowego skupiają się jedynie na kilku obszarach. W literaturze poruszane są również inne problemy, ale do ich rozwiązania są wykorzystywane tradycyjne algorytmy. Pozostawia to pole do wielu potencjalnych usprawnień.

Jednym z poważnych wyzwań w tym obszarze jest ograniczona dostępność sieci QKD. Rozwiązaniem jest zastosowanie symulatorów. Pozwalają one na testowanie działania sieci bez konieczności tworzenia jej fizycznej architektury. W artykule przedstawiono dwa pozwalające na tworzenie złożonych sieci – QKDNetSim oraz QuISP. Są one wciąż rozwijane i w przyszłości powinny pozwolić na przeprowadzenie badań w nietestowanych jeszcze obszarach.

Kwantowa dystrybucja klucza jest przyszłościowym rozwiązaniem. Badania w tym obszarze trwają od wielu lat. Jednak dopiero niedawno zaczęto wykorzystywać w nich uczenie maszynowe. Dotychczasowe wyniki są obiecujące i skłaniają do kontynuacji badań w tym zakresie. Przede wszystkim w kierunkach nakreślonych w niniejszym artykule.

LITERATURA

[1] Rivest, R. L. and Shamir, A. and Adleman, L., "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems", Luty 1978, Association for Computing Machinery, New York, NY, USA

[2] Marcin Niemiec, Andrzej Dziech, Miłosz Stypiński, Jan Derkacz "Quantum-Based Solutions for the Next-Generation Internet" Information & Security: An International Journal, 43 no. 1 (2019): 62-72. https://doi.org/10.11610/isij.4306

[3] ITU-T Y-series Recommendations – Supplement 70 (2021). ITU-T Y.3800-series – Quantum key distribution networks – Applications of machine learning

[4] Chin, HM., Jain, N., Zibar, D. et al. Machine learning aided carrier recovery in continuous-variable quantum key distribution. npj Quantum Inf 7, 20 (2021). https://doi.org/10.1038/s41534-021-00361-x

[5] Zhi-Ping Liu, Min-Gang Zhou, Wen-Bo Liu, Chen-Long Li, Jie Gu, Hua-Lei Yin, and Zeng-Bing Chen, "Automated machine learning for secure key rate in discrete-modulated continuous-variable quantum key distribution," Opt. Express 30, 15024-15036 (2022)

[6] Z. -A. Ren, Y. -P. Chen, J. -Y. Liu, H. -J. Ding and Q. Wang, "Implementation of Machine Learning in Quantum Key Distributions," in IEEE Communications Letters, vol. 25, no. 3, pp. 940-944, March 2021, doi: 10.1109/LCOMM.2020.3040212.

[7] Okey, Ogobuchi Daniel, Siti Sarah Maidin, Renata Lopes Rosa, Waqas Tariq Toor, Dick Carrillo Melgarejo, Lunchakorn Wuttisittikulkij, Muhammad Saadi, and Demóstenes Zegarra Rodríguez. 2022. "Quantum Key Distribution Protocol Selector Based on Machine Learning for Next-Generation Networks" Sustainability 14, no. 23: 15901. https://doi.org/10.3390/su142315901

[8] Y. Cao, Y. Zhao, J. Li, R. Lin, J. Zhang and J. Chen, "Multi-Tenant Provisioning for Quantum Key Distribution Networks With Heuristics and Reinforcement Learning: A Comparative Study," in IEEE Transactions on Network and Service Management, vol. 17, no. 2, pp. 946-957, June 2020, doi: 10.1109/TNSM.2020.2964003.

[9] M. Wenning, S. K. Patri, J. Müller, A. Autenrieth, J. Elbers, P. Rydlichowski, and C. Mas-Machuca, "Towards Optimized Demand Routing in QKD Networks," in Optical Fiber Communication Conference (OFC) 2023, Technical Digest Series (Optica Publishing Group, 2023), paper Th2A.37.

[10] Mehic, M., Maurhart, O., Rass, S. et al. Implementation of quantum key distribution network simulation module in the network simulator NS-3. Quantum Inf Process 16, 253 (2017). https://doi.org/10.1007/s11128-017-1702-z

[11] Quantum Key Distribution Network Simulation Module. Kod źródłowy: https://github.com/QKDNetSim/qkdnetsim-dev (Dostęp maj 2023)

[12] Quantum Key Distribution Network Simulator. https://open-qkd.eu/ (Dostęp maj 2023)

[13] QuISP: a Quantum Internet Simulation Package. / Satoh, Ryosuke; Hajdusek, Michal; Benchasattabuse, Naphan et al. Proceedings - 2022 IEEE International Conference on Quantum Computing and Engineering, QCE 2022. Institute of Electrical and Electronics Engineers Inc., 2022. p. 353-364 (Proceedings - 2022 IEEE International Conference on Quantum Computing and Engineering, QCE 2022).

[14] QuISP: a Quantum Internet Simulation Package. Kod źródłowy: https://github.com/sfc-aqua/quisp (Dostęp maj 2023)

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

Uczenie maszynowe z ochroną prywatności danych

Privacy-preserving Machine Learning

Kamil Sobolak[1]; Piotr Chołda[1]

[1] Akademia Górniczo-Hutnicza im. St. Staszica w Krakowie, Instytut Telekomunikacji, poczta elektroniczna: kamilsobolak@student.agh.edu.pl, piotr.cholda@agh.edu.pl

DOI: 10.15199/59.2023.4.22

Streszczenie: Porównano różne metody służące do zapewniania prywatności w przypadku przetwarzania danych z użyciem uczenia maszynowego. Wybrano najbardziej adekwatne metody: szyfrowanie homomorficzne, prywatność różnicowa, metoda uczenia federacyjnego. Efektywność przedstawionych algorytmów została ujęta ilościowo za pomocą powszechnie używanych metryk: funkcji kosztu dla jakości procesu uczenia, dokładności dla klasyfikacji i współczynnika determinacji dla regresji.

Abstract: Various methods for ensuring privacy in machine learning based data processing were compared. The most suitable methods have been selected: homomorphic encryption, differential privacy, and federated learning. The effectiveness of the presented algorithms was quantified using commonly used metrics: cost function for the quality of the learning process, accuracy for classification, and coefficient of determination for regression.

Słowa kluczowe: prywatność danych, prywatność różnicowa, szyfrowanie homomorficzne, uczenie federacyjne, uczenie maszynowe

Keywords: data privacy, differential privacy, federated learning, homomorphic encryption, machine learning

1. WSTĘP

Uczenie maszynowe wykorzystuje się w celu rozwiązywania zadań klasyfikacyjnych, regresji (prognozowania) i in. Działanie opiera się na udostępnieniu wstępnie wytrenowanych modeli, których celem jest dostarczenie narzędzi wnioskowania, a nie informacji wrażliwych. Ponadto kluczowym elementem algorytmów uczenia maszynowego są dane. W niniejszej pracy porównano trzy metody ochrony prywatności danych współpracujących z uczeniem maszynowym.

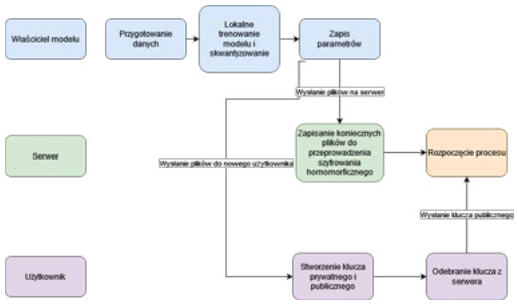
Uczenie federacyjne (ang. *federated learning*) zakłada użycie środowiska składającego się z dużej liczby zdalnych urządzeń i scentralizowanego serwera. Poza tym idea uczenia federacyjnego wpisuje się w paradygmat przetwarzania brzegowego (ang. *edge computing*), polegającego na wykonywaniu obliczeń i przetwarzaniu danych jak najbliżej miejsca ich użycia. Modele przetwarzane są lokalnie i nie ma konieczności ciągłego połączenia z serwerem. Może mieć to kluczowy wpływ na zachowanie prywatności danych, ponieważ eliminuje się główny etap, w którym dane są narażone na ujawnienie i dopasowanie do danego użytkownika (co prowadzi do utraty prywatności). Cały proces składa się z dwóch następujących po sobie kroków: (1) personalizacji danych i (2) agregacji wag serwera [1]. Rozpoczyna się od lokalnego utworzenia modelu i wstępnego procesu uczenia na przechowywanych danych. W następnym kroku dochodzi do wysłania wag na serwer. Potem serwer korzysta z funkcji agregującej, np. średniej ważonej, która bierze pod uwagę liczność zbioru danych użytkownika. W kolejnym kroku serwer rozsyła nowy model bazowy i dochodzi do douczenia przez lokalne urządzenia. Dzięki temu model na danym urządzeniu poprawia się na bieżąco. Wszystkie kroki powtarzane są aż do chwili, gdy model globalny uzyska zadowalające lub ustabilizowane wartości. W publikacjach [1, 2] wprowadzona jest również definicja uczenia federacyjnego. Definiuje się N twórców danych i nazywa się ich klientami $\{C_1, ...C_N\}$. Każdy z nich ma dostęp tylko do swoich danych $\{D_1, ...D_N\}$. W przypadku uczeniu federacyjnego korzysta się z funkcji agregującej wagi, która używana jest na dostarczonych od klientów wagach $\{f_1, ...f_N\}$. Dlatego model scentralizowany nie ma dostępu do zbiorów danych użytkowników.

Prywatność różnicowa (ang. *differential privacy*) jest opisywana jako swego rodzaju obietnica złożona użytkownikowi dostarczającemu dane przez zarządcę bazy danych. Oznacza, że użytkownik nie poniesie żadnych negatywnych skutków po użyciu jego zasobów przez zarządcę danych. Oczekuje się, że jeśli tylko po wykonaniu obliczeń, gdy analizujemy wyniki, nie będzie się dało określić twórcy wejściowego zbioru danych: w takim przypadku istotnie mamy do czynienia z zapewnieniem prywatności różnicowej. Zapewnia to, że indywidualne informacje nie zostaną ujawnione. Metodę stosuje się powszechnie z powodu łatwości wdrożenia, stosunkowo małych wymagań oraz dużej liczby dostępnych implementacji. Opisuje ją wzór [3]:

$$Pr[M(D) \in S] \leq \exp(\epsilon) Pr[M(D') \in S] + \delta \qquad (1)$$

M to losowy algorytm dodający szum Gaussa do danych. S to wszystkie możliwe wyniki, jakie może dać M (te które mogą być przewidziane). Dane deklarowane są na dwa sposoby: D jest to cały zbiór danych, D' to równoległa baza danych, która zazwyczaj nie zawiera wszystkich informacji w porównaniu do poprzedniej (D). Współczynnik ϵ jest definiowany jako maksymalna odległość między dwoma zbiorami. Jest to miara utraty prywatności, zwana budżetem prywatności i definiuje liczbę próbek różniących się między zbiorami. Parametr δ definiuje prawdopodobieństwo wycieku informacji, czyli przypadek, w którym pojedyncza próbka przed i po wykonaniu algorytmu są tożsame.

Ostatnia z metod to **szyfrowanie homomorficzne** (ang. *homomorphic encryption*), które jest metodą zachowania prywatności danych, opierającą się na kryptografii [4]. Protokoły kryptograficzne bardzo często używane są w implementacjach uczenia maszynowego, kiedy wielu użytkowników udostępnia swoje dane. Zapewniają one komunikację z zachowaniem bezpieczeństwa i brakiem wycieku informacji. Nazwa metody składa się z dwóch członów. Pierwszy z nich dotyczy szyfrowania asymetrycznego procesu polegającego na zamianie tekstu jawnego na szyfrogram z użyciem klucza. Homomorfizm opisuje takie transformacje danych, że zachowane są wewnętrzne relacje między nimi. Dzięki temu uzyskujemy dane niepowiązane z żadnym z użytkowników, ale nadające się do użycia w algorytmach uczenia maszynowego. Szyfrowanie homomorficzne pozwala na wykonywania działań matematycznych na szyfrogramie. W wyniku tego, na wejściu algorytmów uczenia maszynowego mamy bezpieczne dane, z których bez odpowiedniego klucza nie jesteśmy w stanie poznać wartości prawdziwych. Wykorzystano dwie metody kwantyzacji, PTQ (ang. *post-training quantization*) [11] i QAT (ang. *quantization aware training*) [12]. Pierwsza z nich używana jest w prostych modelach liniowych (kwantyzacja dodana jest po procesie uczenia na podstawie parametrów wejściowych), natomiast drugą wprowadza się przed szyfrowaniem danych i odpowiedzialność leży po stronie zarządcy danych, aby przekazać odpowiednio przygotowane wartości. Na rysunku 1 przedstawiono w skrócie przebieg procesu zestawiania środowiska z wykorzystaniem szyfrowania homomorficznego. Wynikiem naszego testu jest porównanie predykcji na próbkach użytkownika na zaszyfrowanym modelu w stosunku do uzyskanych bez użycia metody.



Rysunek 1: Schemat działania systemu klient–serwer na potrzeby zastosowania szyfrowania homomorficznego

2. STAN WIEDZY

Prywatność danych jest często poruszana w aktualnych pracach naukowych z obszaru uczenia maszynowego. Wiele z pozycji odnosi się do samej teorii, zasad działania lub przybliża zagadnienia matematyczne kluczowe w zrozumieniu obliczeń. W pracy [5] umieszczono tabelę zbierającą różne zastosowania metody uczenia federacyjnego. W publikacji [6] pokazano użycie metody prywatności różnicowej na zbiorze `MNIST` i `CIFAR-10`. Przeprowadzono obliczenia dla różnej ilości dodanego szumu i zmiennych parametrów δ i ϵ. Zmieniano również takie wartości jak szybkość uczenia, czy liczba warstw sieci neuronowych.

Istnieją metody radzenia sobie ze zbyt mało zróżnicowanymi danymi na wejściu modelu wykorzystującego uczenie federacyjne. Poprawę można uzyskać przez udostępnienie pewnej części danych na samym początku obliczeń i wystartowanie procesu na wstępnie wytrenowanych danych. Dzięki temu już od samego początku użytkownik, który bierze udział w procesie uczenia, może zacząć od wstępnie wytrenowanego modelu. Takie podejście pozwoliło poprawić dokładność modelu o 30% na zbiorze `CIFAR-10` [7]. Inny sposób to odpowiednie powielenie próbek z użyciem tzw. rozszerzania danych (ang. *data augmentation*), np. przycianie obrazów.

Artykuły związane z szyfrowanie homomorficznym w większości skupiają się na implementacji danej metody, warunkach środowiskowych, które muszą być spełnione i na sposobie gromadzenia danych. Dobrym przykładem codziennego użycia jest [8], gdzie autorzy charakteryzują model przewidujący lokalizację. Praca [9] tłumaczy różnicę między implementacjami FHE (ang. *fully homomorphic encryption*) wraz z historią użycia.

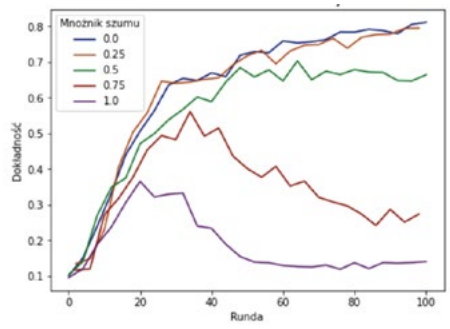
3. IMPLEMENTACJA PORÓWNAWCZA

Celem przedstawionej implementacji jest porównanie sposobów działania różnych metod.

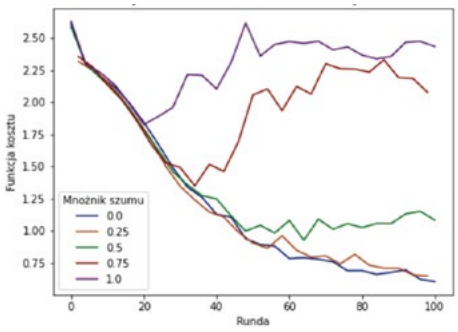
3.1 Uczenie federacyjne i prywatność różnicowa

W tym podrozdziale przedstawiono proces uczenia federacyjnego sieci MLP (ang. *multi layer perceptron*) na zbiorze `EMNIST` dla zmiennej liczby użytkowników w każdej rundzie uczenia i ze względu na mnożnik szumu, który jest miarą odchylenia standardowego szumu Gaussa w stosunku do stałej przełączania. Przed wysłaniem danych na serwer musi zostać dodany odpowiedni szum, tak jak i na serwerze. Jest to zastosowanie prywatności różnicowej. W tym celu użyto metody `FedAvg` (ang. *adaptive clipping*), która dostosowuje udział szumu w każdym z przejść, tak aby jak najlepiej dopasować się do wartości budżetu prywatności. Dynamiczna zmiana parametrów usprawnia działanie algorytmu. Dokładniej cały proces opisany jest w [10], gdzie porównano zmienną i ustaloną wartość stałej przełączania. Podstawę stanowi prosty model trójwarstwowy sieci neuronowej, jednak wymagana była jego adaptacja do użytych dwóch metod ochrony prywatności danych. Uśredniający algorytm federacyjny korzysta z dwóch optymalizatorów: pierwszego dla pojedynczego użytkownika i jego lokalnych aktualizacji modelu. Natomiast drugi optymalizator używany jest dla serwera scentralizowanego. Zaleca się użycie mniejszych wartości szybkości uczenia niż w klasycznym przeprowadzeniu obliczeń. Wiąże się to ze zróżnicowaniem danych wejściowych, może dojść do przypadku kiedy pewne własności zostaną pominięte z powodu losowego dobrania konkretnych użytkowników, których dane nie reprezentują pożądanych cech. Liczba użytkowników w rundzie wpływa na zróżnicowanie danych wejściowych.

Na rysunku 2 pokazano uśrednione wyniki odpowiednio dla 30 serwerów w rundzie z odpowiednio dobranym mnożnikiem szumu tak, aby pokazać, od kiedy wartości dokładności i funkcji kosztu zaczynają odbiegać od pożądanych. Istnieją przypadki, w których dodanie szumu wpływa nieznacznie na wyniki. Wtedy można zauważyć wypłaszczenie charakterystyk w rundach końcowych. Brak wahań krzywej wskazuje na to, że w szczególności model nie jest przetrenowany dla pozytywnych przypadków. Duże anomalie pokazano w sytuacji dodania szumu o większej wartości.



(a) Dokładność



(b) Funkcja kosztu

Rysunek 2: Obliczenia dla 30 serwerów w każdej rundzie

3.2 Uczenie federacyjne z doborem parametru ϵ prywatności różnicowej

Następny rozpatrywany przypadek dotyczy dynamicznego dobrania dodanego szumu i liczby klientów w rundzie w zależności od parametrów prywatności różnicowej dla zbioru `EMNIST`. Jako δ przyjęto 10^{-5}, wynika to z liczby próbek w zbiorze treningowym. Zwyczajowo tę wartość ustawia się na mniejszą niż odwrotność liczby zbioru treningowego. W obliczeniach posługiwano się zmienną wartością ϵ. Zdefiniowano tablicę wejściową składającą się z sześciu wartości od 1,3 do 10 z nierównomiernym skokiem.

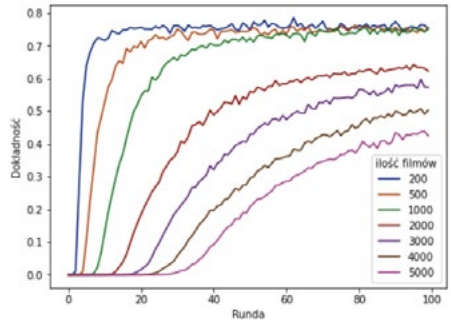
Tabela 1: Wartości uzyskane z parametrami i przedziałami ufności ϵ

ϵ	mn. szu.	liczba s.	funkcja kosztu	dokładność
10	0,5	50	0,6412 ± 0,2122	0,8010 ± 0,0916
5	0,69	69	0,6079 ± 0,1563	0,8102 ± 0,0937
3	0,9	90	0,6155 ± 0,1756	0,8078 ± 0,0874
1.8	1,34	134	0,5913 ± 0,1551	0,8162 ± 0,0794
1.5	1,87	187	0,5898 ± 0,1131	0,8162 ± 0,0731
1.3	3,92	392	0,5956 ± 0,1427	0,8155 ± 0,0714

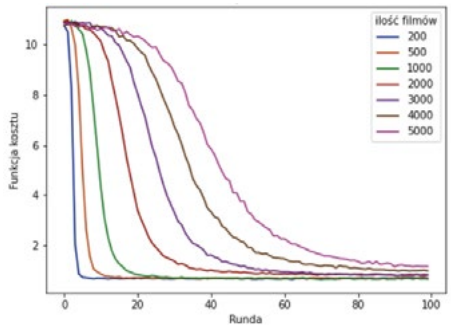
Celem było pokazanie, na ile mała zmiana na pewnych przedziałach wpływa na wyniki. W każdym z przypadków model jest dobrze wytrenowany. Dokładność rośnie, a w końcowych rundach stabilizuje się na poziomie około 0,81 dla każdego przypadku. Natomiast funkcja kosztu maleje i dąży do nieistotnie zmieniającej się wartości w okolicach 0,6. Dowodzi to prawidłowego działanie wyszukiwania parametrów. Wyniki zaprezentowane są w tabeli 1.

3.3 Uczenie federacyjne dla faktoryzacji macierzy

W tym przypadku zaprezentowano technikę faktoryzacji macierzy. Ten matematyczny model, po odpowiednim podziale, umożliwia odnalezienie powiązań i cech wspólnych użytkowników. Jest to metoda tworząca system rekomendacji produktów, zjawisk, elementów na bazie wcześniejszych informacji np. ocen. Proponowanie filmu, korzystając z bazy `Netflix Prize data`, uwzględnia poprzednie oceny użytkownika przez co wpisuję się w ten problem. Dochodzimy do sytuacji, w której mamy n użytkowników, m filmów i tworzona jest macierz o n wierszach i m kolumnach.



(a) Dokładność



(b) Funkcja kosztu

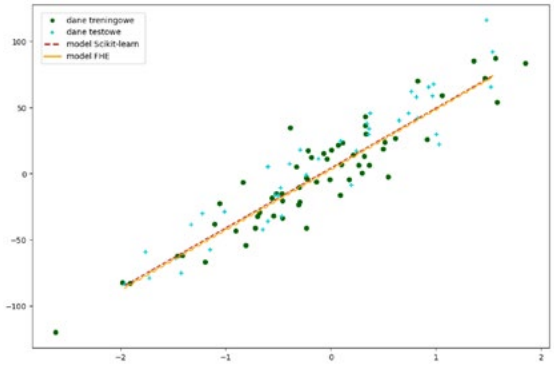
Rysunek 3: Obliczenia uczenia federacyjnego na zbiorze Netflix Prize data

Umieszcza się w niej wszystkie informacje z bazy. W każdym wierszu znajdują się oceny danego użytkownika a każda kolumna zawiera komplet ocen każdego z filmów. Istnieje wiele filmów nie ocenionych przez wszystkich. Podając na wejście algorytmu oceny od jednej osoby możemy przewidzieć jak zostaną sklasyfikowana nieobejrzana pozycja. Dzięki temu uzyskiwana jest informacja, czy dany użytkownik będzie zadowolony po obejrzeniu, co zwiększa chęć korzystania z danego serwisu. Zazwyczaj macierz z poszukiwanymi rekomendacjami jest rzadka, dlatego też sprawdzono, jak bardzo muszą być zróżnicowane dane, aby uzyskać dobry wynik. Na rysunku 3a widać, że wraz ze zwiększoną liczbą filmów w bazie dokładność zmniejsza się nieliniowo. Dla 2000 filmów wartość znacznie się pogarsza. Wpływ na to może mieć mniejsza liczba ocen konkretnego filmu ze zbioru analizowanych danych. Wzrasta prawdopodobieństwo tego, że w dostępnych próbkach nie będzie odnaleziony trend i nie będzie możliwa przyszła ocena, która jest jak najbliższa prawdziwej. W dodatku

szybkość uczenia spada z każdym powiększeniem zbioru, z którego wybieramy oceny. Jeżeli chodzi o rysunek 3b, sytuacja ma się analogicznie jak poprzednio: funkcja kosztu przyjmuje coraz większe wartości przez co spada jakość. Każdy z przypadków w końcowych rundach coraz bardziej się stabilizuje.

3.4 Szyfrowanie homomorficzne z regresją liniową

Ten podrozdział przedstawia wpływ użycia szyfrowania homomorficznego na regresję liniową. Ze względu na ograniczenia FHE należało wprowadzić kwantyzacje danych. Proces ten zmniejsza zróżnicowanie i dokładność danych wejściowych, przykładem tego może być przejście z liczb zmiennoprzecinkowych (ang. *float*) na liczby całkowite (ang. *integer*). Znacznie przyspiesza to czas trwania programu. Miarą jakości jest współczynnik determinacji (ang. R^2 *score*), który określa proporcję wariancji zależnych zmiennych opisanych przez zmienne niezależne. Na rysunku 4 przedstawiono wyniki dla 100 punktów (analogiczne fenomeny można zaobserwować też dla innej liczby). Czerwoną przerywaną linią zaznaczono wynik uzyskany z regresji liniowej (bez ochrony prywatności), natomiast pomarańczową z użyciem FHE. Jak widzimy, linie praktycznie na siebie nachodzą, a nieregularna struktura drugiej z nich związana jest z kwantyzacją.



Rysunek 4: Obliczenia regresji liniowej z szyfrowaniem homomorficznym: 100 punktów

4. PODSUMOWANIE

W niniejszej pracy przedstawiono cztery przykłady, mających na celu zilustrowanie możliwości oraz ograniczeń zaproponowanych w literaturze metod ochrony prywatności danych. Pokazano sposoby użycia wspomnianych metod na danych rzeczywistych i metody zwiększenia poziomu prywatności. W podrozdziale 3.1 wykazano, że można opracować efektywny model używający uczenia federacyjnego i prywatności różnicowej ze zmiennym mnożnikiem szumu i modyfikowaną liczbą klientów w rundzie. Przykład 3.2 obrazuje możliwość zapewnieniami prywatności różnicowej z danymi parametrami poprzez dopasowanie parametrów używanych w obliczeniach (mnożnik szumu i ilość klientów w rundzie). Badano wpływ zmiany ϵ. Przykład 3.3 opisuje udane użycie uczenia federacyjnego wraz z procesem faktoryzacji macierzy. Uzyskano zadowalające wynik dopasowania predykcji oceny filmu na podstawie wcześniejszych ocen użytkowników ze zmieniającą się bazą filmów. W podrozdziale 3.4 zaprezentowano użycie metody szyfrowania homomorficznego w celu zwiększenia poziomu prywatności dla modelu regresji liniowej. Wyniki, po koniecznej kwantyzacji, nieznacznie różniły się od obliczeń bez wprowadzenia metody. Natomiast nie zauważono żadnych zmian między wynikami po kwantyzacji i po kwantyzacji z FHE.

LITERATURA

[1] Dianqi Liu, Liang Bai, Tianyuan Yue, Aiming Zhang. 2022. „Towards Method of Horizontal Federated Learning: A Survey". 8th International Conference on Big Data and Information Analytics (BigDIA)

[2] Qiang Yang, Yang Liu, Tianjian Chen, Yongxin Tong. 2019 . „Federated Machine Learning: Concept and Applications". ACM Transactions on Intelligent Systems and Technology, vol. 10, is. 2, pp 1-19.

[3] Zhanglong Ji, Zachary C. Lipton, Charles Elkan. 2014. „Differential Privacy and Machine Learning: a Survey and Review". CoRR abs/1412.7584 URL: https://arxiv.org/abs/1908.07873 pp. 3-10.

[4] Mohammad Al-Rubaie, J. Morris Chang. 2019. „Privacy-Preserving Machine Learning: Threats and Solutions". IEEE Security & Privacy, vol. 17, is. 2.

[5] Vijaya Kamble, Ashish Phophalia. 2022. „Medical Image Analysis Using Federated Learning Frameworks: Technical Review". IEEE 10th Region 10 Humanitarian Technology Conference.

[6] Martín Abadi, Brendan McMahan, Andy Chu and Ilya Mironov, Li Zhang, Ian Goodfellow, Kunal Talwar. 2016. „Deep Learning with Differential Privacy". ACM SIGSAC Conference on Computer and Communications Security (CCS'16).

[7] Hangyu Zhu, Jinjin Xu, Shiqing Liu and Yaochu Jin. 2021. „Federated learning on non-IID data: A survey". Neurocomputing, vol. 465, is. C, pp 371–390.

[8] Shreya Gupta, Ginni Arora. 2019. „Use of Homomorphic Encryption with GPS in Location Privacy". 4th International Conference on Information Systems and Computer Networks (ISCON).

[9] Chiara Marcolla, Victor Sucasas, Marc Manzano, Riccardo Bassoli. 2022. „Survey on Fully Homomorphic Encryption, Theory, and Applications". Proceedings of the IEEE, vol. 110, no. 10, pp. 1572-1609.

[10] Andrew Galen, Om Thakkar, H. Brendan McMahan, Swaroop Ramaswamy. 2021. „Differentially Private Learning with Adaptive Clipping". Advances in Neural Information Processing Systems vol. 35, pp. 17455-17466.

[11] Qibin Chen, Yijin Teng, Hui Zhang; Kai Jiang, Qiang Duan; Xue Li, Xinxin Zhao, Rui Li. 2022. „Post-Training Quantization for Longformer with Chunkwise Quantization Granularity and Optimized Percentile". 7th International Conference on Computer and Communication Systems (ICCCS).

[12] Jonghoon Kwak, Kyungho Kim, Sang-Seol Lee, Sung-Joon Jang, Jonghee Park. 2022. „Quantization Aware Training with Order Strategy for CNN". IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia).

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

WYKRYWANIE ANOMALII SIECIOWYCH Z OCHRONĄ PRZED ATAKAMI ADWERSARYJNYMI

MACHINE LEARNING BASED NETWORK ANOMALY DETECTION WITH PROTECTION AGAINST ADVERSARIAL ATTACKS

Andres Vejar[1]; Piotr Chołda[2]

Institute of Telecommunications, AGH University of Krakow, Kraków, Poland.
[1]avejar@agh.edu.pl, [2]piotr.cholda@agh.edu.pl

DOI: 10.15199/59.2023.4.23

Streszczenie: **Implementacja sieci neuronowych do wykrywania anomalii sieciowych cieszy się dużym powodzenie. Niestety, tego rodzaju zautomatyzowane metody detekcji są podatne na tzw. ataki adwersaryjne. Atak tego rodzaju polega na spreparowaniu niewielkich odchyleń w stosunku do standardowych danych, co powoduje niepowodzenie w zakresie wykorzystywanego zadania klasyfikacji. Przedstawiamy przegląd metod wykrywania prowadzenia ataków adversaryjnych oraz opisujemy wstępne wyniki uzyskane dzięki zaproponowanej architekturze systemu mającego zapewnić odporność na takie ataki.**

Abstract: **Neural network based implementations for anomaly detection are proven to be successful for a range of communications applications. The inclusion of these automated methods involves vulnerability to adversarial attacks that affect a broad range of models. An adversarial attack consists in fabrication of small variations with respect to the standard data input, with the property of triggering a failure in the classification task at hand. In this study we present a review on adversarial environments, and we describe the preliminary results of our system architecture for the analysis of adversarial-resilient network anomaly detection systems.**

Słowa kluczowe: **Ataki adwersaryjne, Sieci neuronowe, Wykrywanie anomalii sieciowych**

Keywords: **Adversarial attacks, Neural networks, Network anomaly detection**

1. INTRODUCTION

The practicals applications of deep neural network models and their success in outperforming standard algorithms carried out a fast adoption in many fields, including communication networks. Popularized in image recognition models, **adversarial attacks** apply machine learning (ML) inputs that will make those systems fail, for example, an adversarial garment can disrupt a facial camera detection system [12]. The success of those early attacks was driven by the common use of transfer learning to customize image applications from a few original trained image models (e.g. AlexNet, RestNet, VGG-15 YoLo, MobileNet). A collection of trained models is maintained by the Caffe deep learning framework in their Model Zoo repository [5, 7].

Currently, those very large models have increased the computing training cost 300,000 times in six years from 2012, providing very good results, but also exposing a huge parameter space, reaching 175 billion for the Generative Pre-trained Transformer 3 [10], trained with 500 billion tokens. The most extreme case up to date is the PathWays Language Model with 540 billion parameters, pre-trained with a dataset of 780 billion tokens [1]. The computing resources need to be doubled every 6 months to follow the current tendency [9]. Another important aspect of these model behemoths are the availability of training data. Even if the training data is not disclosed, model-inversion attack techniques can estimate and retrieve original samples of the training data [11, 13].

To provide resistance to adversarial attacks, a level of understanding of the underlying model is necessary. We present the following three research directions that can contribute to the safety of the ML models:

- **Adversarial robustness**, where we construct ML classifiers robust to adversarial perturbations on their inputs.
- **Explainable ML**, that enlightens the black-box models of deep learning to construct models that are explainable. Explainable models can be then studied and updated to overcome adversarial attacks.
- **Frugal ML** [4] and **greenAI** [8], that are used to rank the usefulness of models with respect to efficiency and accuracy. Efficient models tend to be shallow, and therefore easy to explain, analyze and secure against adversarial attacks.

More information about anomaly detection in adversarial environments is presented in [2], and in [3] a modern perspective for data generation for network intrusion detection systems is reviewed. For real-time detection, graph based models are used to represent the relationships among the features, and their evolution, see [6] for a recent work considering hypergraphs.

2. ADVERSARIAL ATTACKS

2.1. Classification problem

Adversarial attacks consist in the fabrication of inputs that will allow an attacker to control the output of a supervised classification task. In the context of deep neural network models, the classification problem can be defined as follows:

1. K classification labels: $\text{label}_1, \text{label}_2, \ldots, \text{label}_K$.

2. A dataset $(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)$, where an element (x, y) is composed by the input data $x \in \mathcal{X}$, and its label number $y \in [1, K] \subset \mathbb{Z}$.

3. The classification model $h_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}^K$, where $\boldsymbol{\theta}$ corresponds to the model parameters. $\mathcal{X}$ is the input space and $\mathbb{R}^K$ corresponds to the model solution space, in the form of a membership level vector to the corresponding K classes.

To compare the outputs of $h_{\boldsymbol{\theta}}$, a practical indicator is the probability of membership of the class. To transform the outputs of the model $h_{\boldsymbol{\theta}}$ into probabilities, the softmax operator $\sigma : \mathbb{R}^K \to (0, 1)^K$ is used:

$$\sigma(h_{\boldsymbol{\theta}}(x)) = \frac{\exp[h_{\boldsymbol{\theta}}(x)]}{\mathbf{1}^t \exp[h_{\boldsymbol{\theta}}(x)]}, \tag{1}$$

where $\exp[\cdot]$ corresponds to the entry-wise exponential function. The multiplication of $\mathbf{1}^t$ by Eq. (1), provides a test for the sum of the probabilities, $\mathbf{1}^t \sigma(h_{\boldsymbol{\theta}}(x)) = 1$.

To associate samples with labels in the dataset it is convenient to use a coding matrix $T \in \{0, 1\}^{N \times K}$ with $t_{ik} = 1$ if the sample i belongs to the label k. In a similar way, to evaluate the results of the model with the dataset, a matrix of estimated probabilities $Q_{\boldsymbol{\theta}} \in (0, 1)^{N \times K}$ can be constructed, with rows $q_i = \sigma(h_{\boldsymbol{\theta}}(x_i)), i = 1, \ldots, N$.

Given the coding matrix T, and the estimated probabilities $Q_{\boldsymbol{\theta}}$, the cross-entropy given in Eq. (2) can provide a general estimation of the quality of the model $h_{\boldsymbol{\theta}}$ with respect to the data, accounting for the dissimilarity between $Q_{\boldsymbol{\theta}}$ and T. An objective for the classification is to modify $Q_{\boldsymbol{\theta}}$ to reduce the value in Eq. (2), in consequence approaching the distributions close from the estimate to the coding matrix, as $Q_{\boldsymbol{\theta}} \to T$:

$$H(T, Q_{\boldsymbol{\theta}}) = -\sum_{i=1}^{N} \sum_{k=1}^{K} t_{ik} \log(q_{ik}) \tag{2}$$

To train the classifier, we are required to estimate the parameter vector $\boldsymbol{\theta}$ that minimizes Eq. (2). Setting the loss function $J(\boldsymbol{\theta}) = H(T, Q_{\boldsymbol{\theta}})$, gradient descent-based algorithms will update the parameters using variations of $\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ to minimize the target. Given the scale of the optimization problems, the most used class of algorithms is based in the stochastic gradient descent, that for each iteration partially updates $\boldsymbol{\theta}$ in a subset of directions from $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$.

2.2. Adversarial example attack

Given the characteristics of the classification problem, an adversarial attack with access to a trained model $h_{\boldsymbol{\theta}}$ can modify inputs from $\mathcal{X}$ to achieve miss-classification for the model. The first attack, that we can refer to as adversarial example attack, takes one example input x, estimates its label y with the model, and then it constructs an adversarial example $\hat{x}$ that is near to x and it will trigger failure in the classification task.

A simple approach is to add a small perturbation $\delta \in \Delta$ to the input x, to obtain an adversarial example $\hat{x} = x + \delta$. The set Δ collects the perturbations that will allows to maintain in $\hat{x}$ the true semantic meaning of x. For example, Δ can be the ball of radius ϵ, with $\{\delta : ||\delta||_\infty \leq \epsilon\}$, but more advanced schemes to construct $\hat{x}$ can be considered.

Similarly with Eq. (2), we can use a coding vector $t \in \{0, 1\}^K$, with $t_k = 1$ if the single input x belongs to the label k. For the adversarial example approach, we can maximize the cross-entropy between the estimated value of the input $\hat{x}$ and the true class y, encoded in t:

$$\max_{\hat{x}} H\left(t, q(\hat{x})\right) = \max_{\delta} \left[-\sum_{k=1}^{K} t_k \log\left(q_k(x + \delta)\right) \right], \tag{3}$$

with $q(\hat{x}) = \sigma(h_{\boldsymbol{\theta}}(\hat{x})) \in (0, 1)^K$.

This attack will make the model $h_{\boldsymbol{\theta}}$ fail, and repeating the process for different inputs and labels can provide the attacker with a complete map of adversarial attacks weaknesses of $h_{\boldsymbol{\theta}}$.

2.3. Targeted adversarial attack

The targeted adversarial attack is a directed attack to provide a misclassification falling into a specific label $y^\star$. Following the previous description, we encode the target in the vector $t^\star \in \{0, 1\}^K$, with $t^\star_k = 1$ if the adversarial input $\hat{x} = x + \delta$ belongs to the label k. In this case, we would like to increase the cross-entropy with respect to the true label y and minimize the cross-entropy with respect to the target label $y^\star$:

$$\max_{\delta} \left[H\left(t, q(x + \delta)\right) - H\left(t^\star, q(x + \delta)\right) \right], \tag{4}$$

again with $q(\hat{x}) = \sigma(h_{\boldsymbol{\theta}}(\hat{x})) \in (0, 1)^K$.

The targeted adversarial attack provides an extensive control of the model outputs to the attacker and it is a great concern, specially when pre-trained models are used in critical applications.

3. PROTECTION AGAINST ADVERSARIAL ATTACKS

In Figure 1, we present the architecture of a networked scenario. A parallel system for anomaly detection is deployed to infer the parameters of an adversarial resilient neural network model estimator. The principal defense against adversarial attacks is the construction of models that are aware of adversarial examples. This can be done training the model with a dataset of examples and generate adversarial examples to augment the dataset, therefore the model can learn both types of real examples and adversarial examples. Another possibility is to reduce the original models to explainable models, for example proxy models, regression models and knowledge trees, among others. Finally, the preference of shallow models that are computationally efficient can bypass the use of deep models and therefore, the attacks on shallow models can be easier to manage than in the case of a deep model with hundred billions of parameters. The system of Figure 1 is composed of a LAN communication scenario, with a gateway pipeline that capture, pre-process and extract the features from the network.

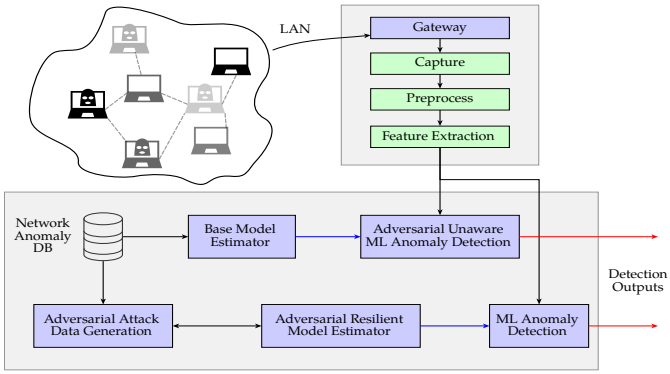


Figure 1: Parallel architecture for adversarial attack protection

The processed data is fed into the detection system, that evaluate an adversarial-unaware anomaly detection model and an adversarial resilient anomaly detection model. The detection outputs are compared to fed back the network anomaly database and the adversarial resilient model estimator. An adversarial attack data generation module is deployed independently of the neural network model for anomaly detection. This choice provides a clear separation of concerns for the different elements of the processing pipeline.

4. IMPLEMENTATION OF THE ATTACKS

In order to test the attacks presented in Section 2, the two cases where studied: (1) Adversarial attacks in image classification using the ResNet50 model. (2) Adversarial attacks in network anomaly detection in the CICIDS2017 dataset using a custom, three-layers deep neural network. The attacks were implemented in a standard computer using 4 cores with `x86_64` architecture and 16 GB RAM running Linux, without GPU acceleration. The algorithms were programmed using the deep learning library PyTorch, version 2.0, compiled natively under Gentoo/Linux, for Python version 3.11. The execution time of the algorithms was limited to few seconds for the attacks, in both cases. Training for the deep neural network for case 2 was limited to 45 minutes.

For implementing the adversarial attack in image classification, we use the pretrained ResNet50 model, that classifies an image with respect to a set of 1000 labels. An instance of the attack is presented in Figure 2, where a photography presenting the Wawel Castle is misclassified as `solar_dish` with probability 0.99 in the adversarial example attack and with the targeted attack to `desktop_computer` it achieves a probability of 0.86.

To test the attacks for network anomalies, the intrusion dataset CICID2017 was used. The original dataset contains 15 labels, considering 14 types of attacks and one 'benign' label. The original data was pre-processed and reduced to 46 features and 5 labels: 'benign', 'botnet', 'brute-force', 'dos ddos', 'infiltration', 'portscan' and 'web attack'. To create a classification model, a multilayer perceptron network was created, with ReLU activation and three layers with sizes 128, 48, and 32, with 14,000 trainable parameters.

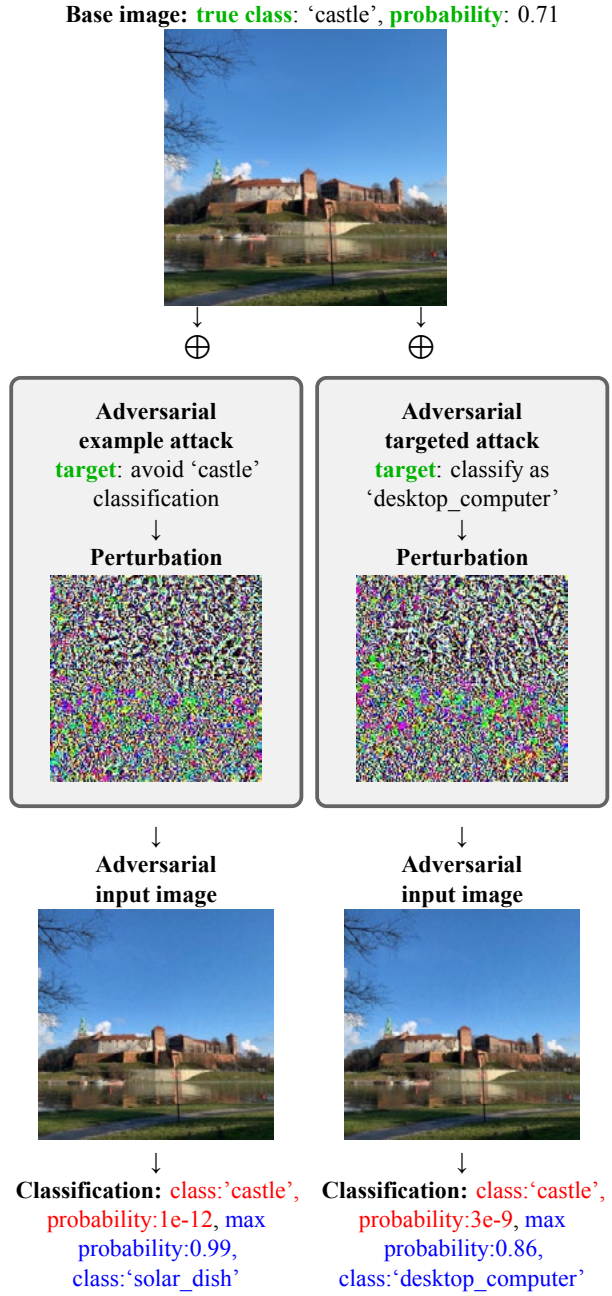


Figure 2: Adversarial attacks to the ResNet-50 model.

The model was trained with 60% of the available data, leaving 20% for model testing, and 20% for model validation. The model achieved correct results with an accuracy of 98.7% on the test dataset. To provide the attack results for the model in a similar way to the previous image case, the 46 features of the model were sorted in the decreasing order with respect to their coefficient of variation ($\hat{\sigma}/\hat{\mu}$) in the original dataset. The features were then mapped to an image of 8x6 pixels, following a counter clockwise spiral matrix from the center of the image to the borders. This way, the highest variance is present in the center and the most static part will go to the borders, providing a stable background. The first two elements of the upper left corner are set in a constant value to account for the 48 pixels correctly. In Figure 3, we present the results for the adversarial attack. Leftmost images are the original inputs x, center images show the perturbation δ, and rightmost images depict the adversarial example $x + \delta$.

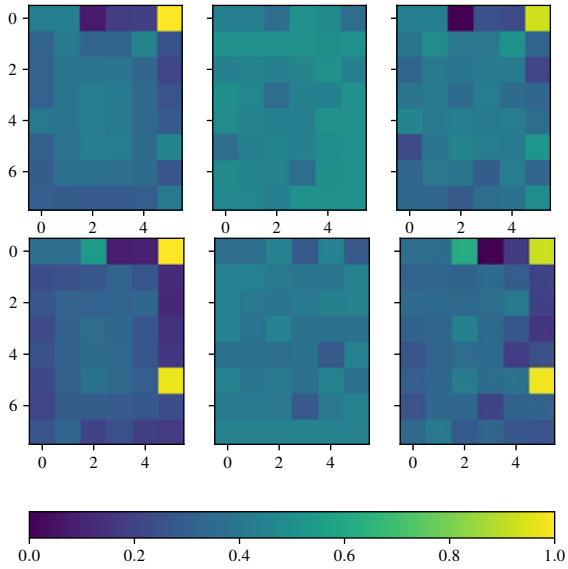


Figure 3: Adversarial attack in the anomaly detection model. The color is mapped to the range of the inputs, from $0 \mapsto \min(x, x+\delta)$, to $1 \mapsto \max(x, x+\delta)$. Pixel coordinates encode feature indexes. Left: original x. Center: perturbation δ. Right: adversarial $x+\delta$. First row 'botnet' to 'benign', second row 'brute-force' to 'benign'.

In the first row, the left image presents the true class 'bot', with an estimated probability of 0.86. After the attack is carried, the model true probability falls to 8.5e-3 and the estimated probability of being 'benign' is 0.99. In the second row, the true class is 'brute-force', with probability 0.99 before the attack. After the attack, the probability of the adversarial example to be in the class 'benign' is 0.98.

5. CONCLUSIONS

In this work, we present the concept of protecting ML based network anomaly detection system against adversarial attacks. To achieve such the protection, the classification models can be trained with adversarial examples, and evaluate the robustness of the generated model with respect to the adversarial examples. Examples of these attacks are provided for image models to illustrate the dangers. An architecture was constructed to show the concept of the system. Using the proposed architecture, the generation of adversarial attacks was studied for a network anomaly dataset. The inputs of the network anomaly model were encoded in images of 8x6 pixels, to provide a simple visualization of the elements of the attacks and the performance of the architecture.

It is important to note that real case scenarios can consider the fine-grain scale of a packet generator to expand the dimension of the input space. That will allow to achieve correct results measured by aggregated metrics, such as *average packet size* or *average flow inter arrival time*. These metrics are considered as features in the CICIDS2017 dataset and in other network intrusion databases. In this preliminary work we have omitted the results presenting the use of training to provide protection against adversarial attacks. Further work will analyze the performance of the adversarial trained model for classification of network anomalies in more detail.

REFERENCES

[1] Anil Ananthaswamy. In AI, is bigger always better? *Nature*, 615(7951):202–205, 2023.

[2] D. Bierbrauer, A. Chang, W. Kritzer, and N. Bastian. Cybersecurity anomaly detection in adversarial environments. *Proceedings of the AAAI Fall 2021 Symposium on AI in Government and Public Sector*, pages 1017—1022, 2021.

[3] Marc Chalé and Nathaniel D. Bastian. Generating realistic cyber data for training and evaluating machine learning classifiers for network intrusion detection systems. *Expert Systems with Applications*, 207:117936, 2022.

[4] Mikhail Evchenko, Joaquin Vanschoren, Holger H. Hoos, Marc Schoenauer, and Michèle Sebag. Frugal machine learning. *CoRR*, abs/2111.03731, 2021.

[5] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, page 675–678, New York, NY, USA, 2014. Association for Computing Machinery.

[6] Zong-Zhi Lin, Thomas D. Pike, Mark M. Bailey, and Nathaniel D. Bastian. A hypergraph-based machine learning ensemble network intrusion detection system. *CoRR*, 2211.03933, 2022.

[7] Berkeley AI Research, The Berkeley Vision and Learning Center, and The Caffe community. Caffe Model Zoo. `https://github.com/BVLC/caffe/wiki/Model-Zoo`, 2020.

[8] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63(12):54–63, 2020.

[9] Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. Compute trends across three areas of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2022.

[10] Tom Brown et al. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[11] Michael Veale, Reuben Binns, and Lilian Edwards. Algorithms that remember: model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180083, 2018.

[12] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V*, page 665–681, Berlin, Heidelberg, 2020. Springer-Verlag.

[13] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 253–261, 2020.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

ANALIZA WSKAŹNIKA φ POD KĄTEM PRZYDATNOŚCI JAKO MIARY W OCENIE ZŁOŻONOŚCI SZYFROWANIA I JAKOŚCI SZYFROGRAMU W METODZIE S-TECH

mgr Przemysław Bezeg[1]; dr hab. inż. Piotr Pacyna, prof. AGH[2]

[1] Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie,
Wydział Informatyki, Elektroniki i Telekomunikacji Kraków, kontakt: przemyslaw.bezeg@gmail.com

[2] Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie,
Wydział Informatyki, Elektroniki i Telekomunikacji Kraków, kontakt: piotr.pacyna@agh.edu.pl

DOI: 10.15199/59.2023.4.24

Streszczenie: **Artykuł dotyczy szczególnego rodzaju szyfrowania wiadomości, któremu towarzyszy ukrywanie szyfrogramu pod postacią tekstu. W efekcie otrzymujemy szyfrogram w formie tekstu, który jest poprawny stylistycznie i semantycznie, a więc zbliżony do tekstu naturalnego. W toku badań analizujemy metodę szyfrująco-ukrywającą s-Tech, a w szczególności jej wskaźnik φ, który służy do oceny trudności generowania szyfrogramu i do szacowania jakości wynikowego tekstu, to jeststopnia naturalizmu powstającego szyfrogramu. Celem badań jest sprawdzenie użyteczności tej miary jako uniwersalnego wskaźnika złożoności przebiegu szyfrowania i jakości tekstu. Badanie wskaźnika φ odbywa się poprzez manipulację dwoma parametrami systemu: długością n-Gramów w bazie n-Gramowej (w zakresie od *n*=1 do *n*=6, oznaczanej też jako LBS) oraz włączając (albo wyłączając) przetwarzanie wstępne. Oceniamy ich łączny wpływ – nie tylko na przebieg szyfrowania, na trudność, lecz również na jakość szyfrogramu. Analiza odbywa się poprzez porównanie wyników dla trzech wariantów preprocessingu: szyfrowanie hybrydowe połączone z kompresją LZW, kompresja SMAZ oraz dla sytuacji referencyjnej, w której tekst jawny w zapisie ASCII jest szyfrowany bez przetwarzania wstępnego.**

Abstract: **The paper focuses on a unique encryption method combined with shaping ciphertext as natural text, which is a form of steganography. We analyze the s-Tech encryption method and its φ indicator by evaluating the difficulty of ciphertext generation and the quality of the resulting natural text. The research aims to examine φ as a universal indicator of both encryption complexity and natext quality. The analysis involves three preprocessing variants: hybrid encryption with LZW compression, SMAZ compression, and a reference situation with null preprocessing.**

Słowa kluczowe: **NLP, n-Gram, Łańcuchy Markova, Szyfrowanie.**

1. WSTĘP

Metoda s-Tech została opracowana w celu wprowadzenia mechanizmu zawierającego funkcję szyfrująco-ukrywającą dane poprzez generowanie tekstów wynikowych, które są zbliżone do takich, które mógłby utworzyć człowiek. Jej zasadniczym elementem jest mechanizm jednocześnie szyfrujący wejściowe dane i ukrywający powstający szyfrogram. Przyjmuje on dane wejściowe w postaci bloków i przekształca je na tekst, który ma naturalną postać i jest częściowo zrozumiały dla postronnego czytelnika, a jednocześnie nie ujawnia treści szyfrogramu ani jego obecności. Powstający dynamicznie szyfrogram przyjmuje postać tekstu, który nie odbiega leksykalnie i stylistycznie od tekstu utworzonego przez człowieka, choć posiada pewne braki w warstwie semantycznej.

Omawiana metoda wykorzystuje szyfr blokowy, który wykorzystuje weryfikator poprawności w postaci baz n-Gramowych, zawierających 2 018 754 509 *n*-Gramów o długości do 10 wyrazów. Wykorzystanie baz *n*-Gramowych o różnych wartościach *n*, w przedziale <1,6>, ma istotny wpływ na semantyczną spójność generowanego tekstu wynikowego (szyfrogramu), a więc na jakość.

Do stworzenia n-Gramowej bazy danych użyto korpusu Wikipedii w języku angielskim (Wikipedia, 2021), co pozwoliło na syntetyczne generowanie szyfrogramów w postaci tekstów strukturalnie i semantycznie zbliżonych do takich, jakie tam występują. Teksty w Wikipedii posiadają obszerną treść, są różnorodne tematycznie, bogate w kontekst oraz są napisane przystępnym językiem. Te cechy mają na celu zapewnienie, że generowany tekst będzie naturalny, zaś ukryty przekaz będzie trudny do wykrycia oraz niemożliwy do nieuprawnionego odszyfrowania.

Ochrona prywatności: Metoda s-Tech może być wykorzystana do ochrony prywatności użytkowników w sieciach społecznościowych i systemach telekomunikacyjnych. Poprzez zastosowanie technik homofonicznych i przetwarzania danych, metoda s-Tech może zapewnić anonimowość i poufność danych osobowych, co jest szczególnie istotne w przypadku transmisji wrażliwych informacji.

W zastosowanej metodzie ukrywania danych wykorzystano techniki NLP (*Natural Language Processing*), jak również Łańcuchy Markowa oraz przetwarzanie dużych zbiorów danych (*Big Data*).

Warto tutaj nadmienić, że mimo rozwoju metod przetwarzania języka naturalnego (NLP) obecnie nie istnieje w pełni skuteczna metoda automatycznego rozpoznawania bez udziału człowieka, czy dany tekst został stworzony przez człowieka, a kiedy jest naturalny i prawidłowy, to czy zawiera ukryty przekaz. Zdanie "Bezbarwne zielone idee wściekle śpią" (*ang.* Colorless green ideas sleep furiously (Chomsky 1956), jest gramatycznie poprawne, jednak zupełnie pozbawione sensu. Potrafimy generować tekst i możemy próbować ocenić, stosując różne techniki czy zdanie jest poprawne gramatycznie,

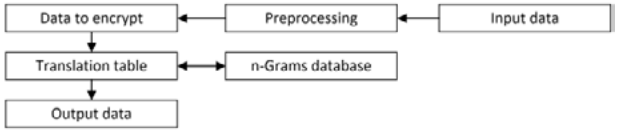
stylistycznie i znaczeniowo, ale żadna z dotychczas znanych nie gwarantuje, że wyprodukowany tekst będzie osiągał poziom zbliżony do tekstu pochodzącego od człowieka. Obecne algorytmy nie są w stanie poradzić sobie z takimi zadaniami, choć notowany jest szybki postęp na tym polu. Oczywiście, jeśli rozpatrujemy zdanie w kontekście poetyckim, możemy dokonywać interpretacji i doszukiwać się w nim pewnego przesłania, co jednak jest subiektywne i niezwiązane z postawionym tutaj przeznaczeniem tekstu (jako nośnika szyfrogramu), który spełnia wymaganie odwracalności operacji - ekstrakcji tekstu jawnego. Dlatego w niniejszym artykule nie chodzi wyłącznie o wysoką jakość tekstu, lecz o jego funkcję jako nośnika ukrytego przekazu. Wykrycie występowania w nim szyfrogramu, kiedy tekst jest naturalny, również winno być trudne albo niemożliwe.

We wcześniejszych badaniach, nieujętych w tym artykule, porównano jakości różnych tekstów wynikowych z wielu przebiegów szyfrowania poprzez ocenę ekspercką. Wytypowano wówczas wskaźnik φ, ponieważ jego wartości były dodatnio skorelowane z jakością tekstu. Pozwala to uznać wskaźnik φ równocześnie jako indykator jakości tekstu jak i trudności jego generowania.

Oprócz metody s-Tech, istnieje jeszcze metoda NICETEXT (Chapman 1997), która na wyjściu podaje również tekst ukryty a także RNN-Stega (Z.-L. Yang 2019) która operuje na sieciach neuronowych.

2. OPIS METODY

Omawiany system szyfrująco-ukrywający łączy kilka metod, które umożliwiają automatyczne generowanie tekstu. Na (*Rys. 1.*) przedstawiono schemat metody.

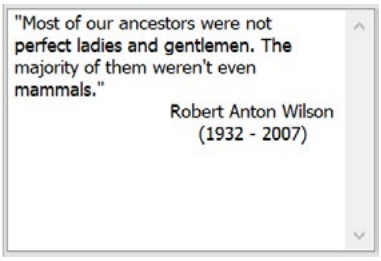


Rys. 1. Ogólny wykres blokowy metody s-Tech

Metoda s-Tech szyfruje i ukrywa tekst jawny (dane), które mogą być poddane preprocessingowi, na przykład kompresji. Następnie bloki danych wejściowych sa przekształcane według tablicy translacyjnej na wyrazy i łączone z kolejnymi zamienionymi w taki sposób wyrazami tak, aby utworzyły ciąg występujący w n-Gramowej bazie danych. Podobnie pobierane są kolejne bloki i zamieniane wg tablicy translacyjnej na kolejne wyrazy celem utworzenia ciągu słów znanego z n-Gramowej bazy danych. W przypadku utworzenia ciągu słownego, który nie występuje w n-Gramowej bazie danych, usuwany jest ostatni wyraz powstającego tekstu wyjściowego, algorytm wykonuje jeden krok wstecz i próbuje „dobrać" inne słowo celem utworzenia pełnego ciągu, występującego w n-Gramowej bazie danych. Słowa „odrzucone" implikują kroki wstecz, które są zliczane jako niepowodzenia i na samym końcu szyfrowania prezentowane jako wskaźnik *φ*.

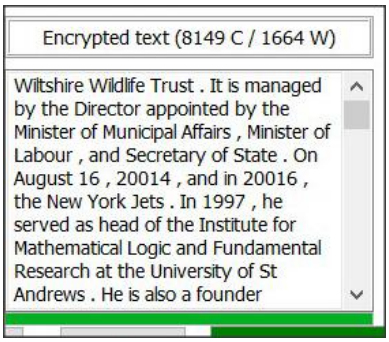
3. PRZYKŁAD

Rys. 2 prezentuje tekst jawny (*plaintext*) o długości 208 bajtów, gdzie ostatnie dwie linie zostały poprzedzone spacjami (27 i 32 spacje) w celu zbadania wpływu przetwarzania wstępnego (*preprocessingu*) takiego jak kompresja lub szyfrowanie, na wartości przyjmowane przez wskaźnik φ. Rys. 3 zawiera tekst po zaszyfrowaniu.


"Most of our ancestors were not perfect ladies and gentlemen. The majority of them weren't even mammals."

Robert Anton Wilson
(1932 - 2007)

Rys. 2. Plaintekst


Encrypted text (8149 C / 1664 W)

Wiltshire Wildlife Trust . It is managed by the Director appointed by the Minister of Municipal Affairs , Minister of Labour , and Secretary of State . On August 16 , 20014 , and in 20016 , the New York Jets . In 1997 , he served as head of the Institute for Mathematical Logic and Fundamental Research at the University of St Andrews . He is also a founder

Rys. 3. Szyfrogram wytworzony metodą s-Tech

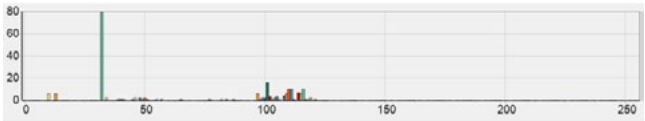
4. PRZETWARZANIE WSTĘPNE ORAZ PARAMETR φ

W języku angielskim występuje 26 liter, przedstawianych za pośrednictwem 52 znaków ASCII, a dodatkowo 10 cyfr oraz znaki przestankowe i specjalne. Reprezentacja jako kody ASCII prowadzi do skupienia wokół pewnych wartości liczbowych. Różnice w częstościach występowania są pogłębione przez zróżnicowaną częstość występowania pewnych znaków w tekście naturalnym.

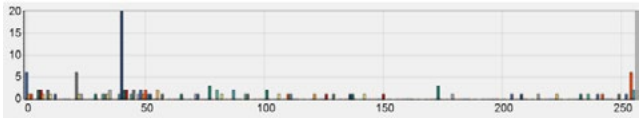
Na *Rys. 2* przedstawiono tekst jawny, który zostanie wykorzystany jako materiał wsadowy. W dwóch ostatnich liniach dodano odpowiednio 27 i 32 znaki spacji w celu zbadania ich wpływu na generowany tekst wynikowy w porównaniu do zastosowania kompresji SMAZ (Sanfilippo 2020) i szyfrowania hybrydowego. Trzy metody preprocessingu mają na celu rozproszenie kodów znaków w pełnym zakresie wartości, przy bloku jednostkowym o wielkości 8 bitów.

Rys. 4, *Rys. 5 i Rys. 6* przedstawiają częstości występowania wartości kodów ASCII cytatu z *Rys. 2*: brak preprocessingu (*Rys. 4*), kompresja SMAZ (*Rys. 5*) oraz szyfrowanie hybrydowe (*Rys. 6*).

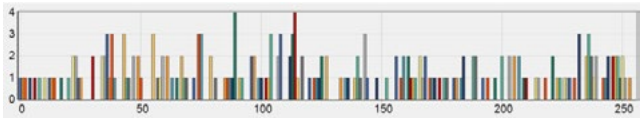
Osie rzędnych reprezentują kod znaku w notacji ASCII w zakresie wartości 0-255. Osie odciętych wskazują liczbę wystąpień znaku o konkretnym kodzie ASCII. Przeważają częstości wystąpień mieszczące się w przedziale od 0 do 16 (*Rys. 4*), w przedziale 0 do 6 (*Rys. 5*) i w przedziale 0 do 3 (*Rys. 6*). Istotną różnicą widoczną na wykresach jest koncentracja wystąpień lub jej brak wokół pewnych wartości. Zauważalna jest również dominanta wynikająca z zastąpienia białego znaku (spacji) w tekście na inny znak przy kompresji SMAZ.



Rys. 4. Częstości występowania wartości kodów ASCII dla cytatu z (Rys. 2) w bajcie bez preprocessingu



Rys. 5. Częstości występowania symboli cytatu z (Rys. 2) - SMAZ



Rys. 6. Częstości występowania symboli cytatu z (Rys. 2) - szyfrowanie hybrydowe

Najlepsze rozproszenie występuje wskutek użycia Szyfrowania Hybrydowego (J.H. Urbanik 2015), podczas gdy najgorsze przy braku szyfrowania (por. *Rys. 4 ÷ Rys. 6*). Obserwując powyższe wykresy częstości występowania symboli można przypuszczać z samej konstrukcji funkcji szyfrującej, że rozproszenie wartości symboli będzie miało wpływ na wartości osiągane przez wskaźnik φ, co powinno być widoczne w dalszych testach.

Przetwarzanie wstępne poprzez kompresję tekstu jawnego (*Rys*. 2) metodą SMAZ zmniejszyło jego wielkość z 208 do 117 bajtów. Dlatego, aby można było porównać tą metodę preprocessingu z pozostałymi dwiema, dane uzyskane podczas prowadzonych w dalszej części szyfrowań zostały przeskalowane przez współczynnik 1.77. W przypadku szyfrowania hybrydowego uzyskano 208 bajtów, co nie wymaga skalowania wyników.

Badając 6 baz n-Gramowych utworzonych z korpusu Wikipedii, każda zawierająca ciągi słów (n-Gramy o innej długości, w zakresie od n=1 do n=6) oraz trzy rodzaje preprocessingu, otrzymujemy łącznie 18 eksperymentów. W ramach każdego, poprzez szyfrowanie tego samego ciągu tekstowego (*Rys. 2*) zbadamy wartości wskaźnika φ . Należy podkreślić, że metoda s-Tech generuje różne teksty wyjściowe dla każdego wykonania (przebiegu, iteracji), co oznacza, że uzyskane wartości wskaźnika φ będą inne w każdym przebiegu (iteracji), mimo operowania wciąż na tym samym tekście wejściowym. Dla uściślenia, każdy z 18 rodzajów testów zostanie powtórzony 10 000 razy (iteracji), a osiągane wartości φ zostaną przedstawione w formie histogramów.

Wartość 10 000 iteracji została wybrana aby utworzyć histogramy z większą dokładnością w sytuacji, kiedy mamy dużą liczbę przedziałów i wobec tego małą liczbę zliczeń w każdym przedziale. W przypadku mniejszej liczby iteracji, wskazania obrazowane na histogramach wykazywały znacznie większe zróżnicowanie wartości w sąsiadujących przedziałach i w konsekwencji nie ukazywały wykresu zbliżonego do krzywej Gaussa.

Prawe końce histogramów zostały w celach prezentacji skrócone (wszystkie przedziały o licznościach mniejszych niż 1% wartości mediany zostały odcięte). Suma wszystkich odciętych wartości została przedstawiona jako dodatkowy element wykresu na prawych krańcach histogramów.

5. WSKAŹNIK φ

Wskaźnik φ stanowi wytypowany przez nas licznik niepowodzeń, które mogą wystąpić podczas kompletowania kolejnego (kandydującego) wyrazu, rozwijającego bieżący ciąg symboli (słów) szyfrogramu, o nowe słowo. Niepowodzenia są istotne, gdyż wskazują na trudności w konstruowaniu szyfrogramu o wymaganej długości frazy, na przykład n=6. Po wybraniu nowego, kandydującego wyrazu następuje sprawdzenie, czy nowopowstający ciąg wyrazów, z powiększenia dotychczasowego o kolejny wyraz (następnik) znajduje się w bazie n-Gramów, a więc pośrednio, czy występuje w korpusie. Jeżeli nowoutworzony ciąg wyrazów występuje w bazie, wówczas algorytm kontynuuje działanie bez zwiększania wartości wskaźnika φ, a w przeciwnym razie odnotowuje niepowodzenie zwiększając wartość φ o 1 i podejmuje nową próbę: przystępuje do wyboru alternatywnego wyrazu, który podobnie jak ten odrzucony staje się kandydującym wyrazem wydłużającym ciąg słów szyfrogramu. Losowanie nowego wyrazu odbywa się w oparciu o wartość szyfrowanego elementu wejściowego z pewnego podzbioru dostępnych równoważnych słów.

6. BADANIE WSKAŹNIKA φ

Po przeprowadzeniu 18 serii badań (3 metody przetwarzania wstępnego, 6 poziomów głębokości lub inaczej dopuszczalnej liczby kroków wstecz), otrzymano wyniki przedstawione w (*Tab. 1*). Tabela zawiera mediany wartości wskaźnika φ, uzyskane w seriach. Z kolei (*Tab.2*) zawiera sumy wartości wskaźnika φ z wszystkich iteracji w obrębie serii. Tabele median i sum zostały opracowane według długości bazy n-Gramów (opisanej jako LBS).

Tab. 1 Mediany wskaźnika φ

MEDIANS	Hybrid	SMAZ	None
LBS = 1	34	32	34
LBS = 2	272	284	390
LBS = 3	1 437	1 483	1 520
LBS = 4	5 249	5 463	5 590
LBS = 5	22 208	23 830	24 058
LBS = 6	127 332	147 615	127 278

Tab. 2. Sumy wskaźnika φ

INTEGRALS	Hybrid	SMAZ	None
LBS = 1	10 000	10 000	10 000
LBS = 2	60 000	90 000	120 000
LBS = 3	200 000	270 000	220 000
LBS = 4	780 000	1 160 000	800 000
LBS = 5	4 390 000	8 710 000	5 010 000
LBS = 6	38 370 000	148 090 000	80 140 000

Mediany wartości wskaźnika φ (*Tab. 1*) ilustrują, że wraz z poprawą losowości danych wejściowych (None, SMAZ, Hybrid) obserwuje się, że wartość wskaźnika φ maleje, co oznacza, że szyfrogram powstaje „łatwiej" (przy mniejszej liczbie niepowodzeń), co przekłada się na potencjalnie lepszą jakość tekstu wynikowego.

Należy też nadmienić, czego nie możemy pokazać z racji ograniczonej wielkości artykułu, że zwiększanie wartości LBS (długość ciągu n-Gramowego) przyczynia się do poprawy jakości tekstu wynikowego.

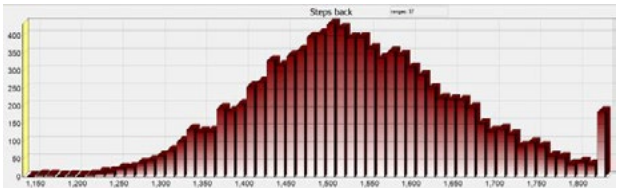
W ramach każdego przedziału można określić współczynnik oczekiwanej jakości tekstu.

Na przykład, histogram częstości przyjmowania przez wskaźnik φ różnych wartości w iteracjach, dla serii LBS = 3 przy braku preprocessingu jest przedstawiony na (*Rys. 7*). Ostatni przedział na wykresie reprezentuje zbiorczą liczbę elementów pominiętych na histogramie.

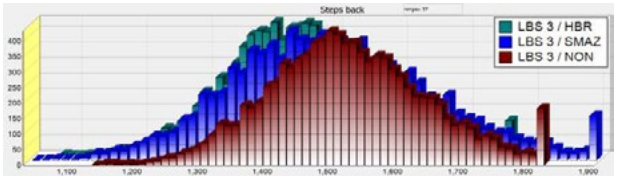
Kolejny rysunek (*Rys. 8*) przedstawia wzajemne przesunięcie histogramów wartości parametrów φ, co ilustruje pochodzenie różnic pomiędzy wartościami median (z *Tab. 1)* w zależności od rodzaju preprocessingu – w nawiązaniu do (*Rys. 4*), (*Rys. 5*) oraz (*Rys. 6*). Z kolei na histogramach na (*Rys. 9*) można zaobserwować odległe przesunięcia całych grup histogramów wartości wskaźnika φ z uwagi na różne długości kompletowanych n-Gramów (różne wartości LBS) oraz zmiany kształtu histogramów. Te odległości zostały zobrazowane na wykresie (*Rys. 10*), gdzie oś odciętych przyjmuje skalę logarytmiczną i który obrazuje wartości z (*Tab. 1)*. Obserwujemy wykładniczy wzrost trudności utworzenia szyfrogramu w funkcji długości n-Gramów.

Osie rzędnych na wykresach (*Rys. 8-10*) oznaczają zakresy dla przedziałów wartości φ, które zostały zsumowane w iteracjach.

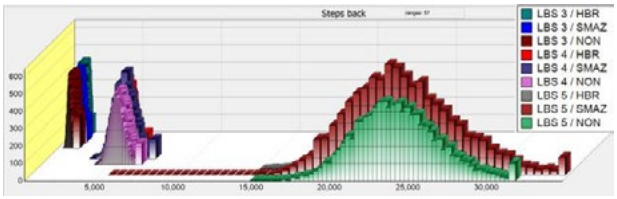
Sumy wartości wskaźnika φ, zliczone ze wszystkich iteracji, zostały przedstawione dla różnych serii w (*Tab.2*) i zilustrowane na wykresie (*Rys. 11*), gdzie oś odciętych również ma skalę logarytmiczną.



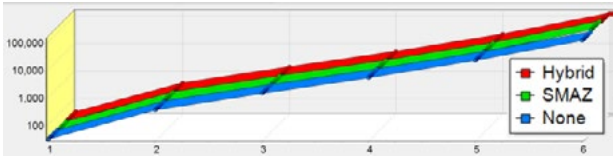
Rys. 7 Histogram częstości wskaźnika φ – seria LBS=3, None.



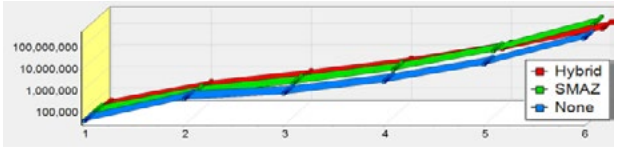
Rys. 8. Częstości φ – seria dla LBS=3 (None, SMAZ, Hybrid)



Rys. 9. 6 histogramów dla LBS=3, 4, 5



Rys. 10. Mediany wartości wskaźnika φ w funkcji liczby kroków wstecz LBS (por. Tab.2)



Rys. 11. Sumy wartości wskaźnika φ ze wszystkich iteracji w funkcji liczby kroków wstecz LBS (por. Tab. 2)

7. WNIOSKI

Wskaźnik φ jest istotnym wskaźnikiem jakości wygenerowanego tekstu w metodzie s-Tech, jednak wyciąganie wniosków w oparciu o jego wartość powinno być rozpatrywane oddzielnie dla różnych długości n-Gramu (wartości LBS). Każde zwiększenie LBS skutkuje znaczącym wzrostem wartości wskaźnika φ, co także wpływa pozytywnie na jakość tekstu wynikowego. Zatem porównywanie wartości φ pochodzących z szyfrowania przy różnych wartościach LBS jest niewłaściwe, bo może prowadzić do nieprawidłowych wniosków.

Podsumowując (1) Wskaźnik φ dostarcza informacji na temat jakości danych wygenerowanych przez metodę s-Tech, w tym dla różnych metod preprocessingu, (2) nie należy porównywać wartości φ jeśli pochodzą z szyfrowania przy różnych długościach n-Gramu (różnych wartościach LBS).

Zauważono nieoczekiwany spadek wartości φ przy LBS=6 (bez przetwarzania wstępnego) - zarówno mediany jak i sumy wartości wskaźnika φ ze wszystkich iteracji. Jest to przypuszczalnie związane z naturalną i powtarzalną strukturą gramatyczną zdań w języku angielskim oraz statystyczną długością zdań w korpusie Wikipedii. Ta sytuacja wymaga dalszej analizy i interpretacji.

Pragniemy gorąco podziękować anonimowym recenzentom za cenne uwagi oraz ich wkład do poprawy jakości niniejszego artykułu.

LITERATURA

[1] Chapman M. T., George I. Davida. 1997. *Hiding the hidden: a software system for concealing ciphertext as innoculous text.* University of Wisconsin-Milwaukee: springer ICICS 1997: Information and Communications Security pp 335-345.

[2] J.H. Urbanik Bezeg Przemyslaw. 2015. Method and system for a symmetric block cipher using a plurality of symmetric algorithms. USA Patent US 8942371 B2.

[3] Noam Chomsky. 1956. *Three models for the description of language.* Massachusetts: Department of Modern Languages and Research Laboratory of Electronics Massachusetts Institute of Technology Cambridge.

[4] S. Sanfilippo. 2020. *Github SMAZ.* Data uzyskania dostępu: 07 01, 2020. https://github.com/antirez/smaz.

[5] Y.-J. Z.-L. Yang, Y.-F. Huang, Z.-M. Chen, X.-Q. Guo 2019. „RNN-Stega: Linguistic steganography based on recurrent neural networks.” *IEEE Transactions on Information Forensics and Security* vol. 14 (Issue: 5, May 2019). doi:https://doi.org/10.1109/TIFS.2018.2871746.

WYSOCE BEZPIECZNA ŁĄCZNOŚĆ SPECJALNA NA BAZIE NOWOCZESNYCH RADIOSTACJI WOJSKOWYCH

HIGH SECURE COMMUNICATIONS BASED ON MODERN TACTICAL RADIO

Jarosław Wojtuń[1]; Kacper Bednarz [2]; Krzysztof Maślanka[3];

[1] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, jaroslaw.wojtun@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, kacper.bednarz@wat.edu.pl
[3] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, krzysztof.maslanka@wat.edu.pl

DOI: 10.15199/59.2023.4.25

Streszczenie: **Bezpieczna wymiana informacji ma w obecnie wykorzystywanych systemach ogromne znaczenie. Jeszcze większego znaczenia nabiera w przypadku systemów przetwarzających informacje wrażliwe, które funkcjonują najczęściej jako systemy wydzielone z odpowiednio przygotowaną infrastrukturą przewodową. Jednocześnie Resort Obrony Narodowej wyposaża Siły Zbrojne Rzeczypospolitej Polskiej w nowoczesne radiostację, które są certyfikowane do przetwarzania i przesyłania informacji o najwyższych klauzulach tajności. W artykule zaproponowano koncepcję wykorzystania szyfratorów tych radiostacji do zapewnienia bezpiecznego kanału transmisyjnego przez publiczną sieć telekomunikacyjną, również bezprzewodową. Przedstawiona architektura, jak wykazały testy, nie wpływa znacząco na pogorszenie jakości realizowanych usług, zapewniając wysokie bezpieczeństwo przesyłanych informacji.**

Abstract: **The secure exchange of information is of great importance in the systems currently in use. It becomes even more important in the case of systems that process sensitive information, which usually function as separated systems with a properly prepared wired infrastructure. At the same time, the ministry of defence equips the Army with modern radios that are certified to process and transmit information with the highest security clauses. The article proposes the concept of using the encryption devices of these radio to provide a secure transmission channel over the public telecommunications network, including wireless. The presented architecture, as tests have shown, does not significantly affect the deterioration of the quality of service provided, ensuring high security of the transmitted information.**

Słowa kluczowe: **AN/PRC-117G, bezpieczeństwo, HAIPE, IPSec, LTE.**

Keywords: **AN/PRC-117G, security, HAIPE, IPSec, LTE.**

1. WSTĘP

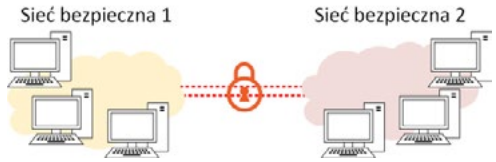
Informacja jest bardzo ważnym czynnikiem pozwalającym osiągnąć zamierzony cel np. polityczny, gospodarczy, finansowy. Bezpieczeństwo informacji to szereg przedsięwzięć, które mają na celu chronić informację danego przedsiębiorstwa, instytucji czy państwa przed nadużyciami, nieautoryzowanym dostępem, zakłóceniami lub zniszczeniem. Zgodnie z obowiązującymi w Polsce aktami prawnymi „bezpieczeństwo informacji niejawnych przetwarzanych w systemie teleinformatycznym zapewnia się przez wdrożenie spójnego zbioru zabezpieczeń w celu zapewnienia poufności, integralności i dostępności tych informacji" [12, 14].

Polityka bezpieczeństwa Resortu Obrony Narodowej wymaga równoległego realizowania procesu organizacji ochrony sieci teleinformatycznych i zapewnienia bezpieczeństwa przesyłanej informacji na wszystkich etapach planowania. Jest to szczególnie istotne w dobie gwałtownego rozwoju technologicznego. Nieustannie rośnie zapotrzebowanie na coraz bezpieczniejsze metody i środki do przekazywania informacji.

Na system teleinformatyczny oraz przetwarzaną w nim informację wpływa szereg bezpośrednich zagrożeń. Zaliczają się do nich błędy personelu, zła organizacja pracy, awarie i uszkodzenia sprzętu, czy też siły wyższe. Największym zagrożeniem są jednak celowe działania osób trzecich, nierzadko całych organizacji, których celem jest przechwycenie informacji wykorzystując przy tym wszelkie możliwe środki [8].

Wymagania bezpieczeństwa informacji stawiane sieciom teleinformatycznym infrastruktury krytycznej państwa oraz sieciom wojskowym obejmują zagadnienia związane z bezpieczeństwem osobowym, bezpieczeństwem źródeł informacji oraz kontrolą dostępu do zasobów sieci teleinformatycznej. System teleinformatyczny, w którym mają być przetwarzane informacje wrażliwe, w szczególności niejawne poddaje się akredytacji bezpieczeństwa teleinformatycznego. W Polsce podmiotami, które udzielają akredytacji bezpieczeństwa są Agencja Bezpieczeństwa Wewnętrznego i Służba Kontrwywiadu Wojskowego. Określone są wytyczne jakie powinien spełniać system teleinformatyczny [1, 13].

Sytuacja komplikuje się w przypadku konieczności wymiany danych pomiędzy dwoma lub więcej systemami (sieciami teleinformatycznymi), w szczególności, gdy systemy te znajdują się w różnych, często odległych lokalizacjach. Scenariusz taki, w sposób poglądowy, obrazuje rys. 1.



Rys. 1. Koncepcja bezpiecznego połączenia dwóch sieci teleinformatycznych

Rysunek przedstawia dwie sieci, spełniające określone wymagania bezpieczeństwa, które czynią je sieciami bezpiecznymi. W celu dystrybucji danych pomiędzy tymi sieciami konieczne jest zbudowanie bezpiecznego połączenia. Efekt ten można osiągnąć poprzez stworzenie własnego bezpiecznego kanału łączności, jednakże jest to rozwiązanie kosztowne i czasochłonne. Alternatywnym rozwiązaniem jest wykorzystanie istniejącej infrastruktury telekomunikacyjnej. Bezpieczeństwo transmisji danych w tym wariancie zapewnia się poprzez zastosowanie odpowiednich rozwiązań sprzętowo-programowych.

Powszechnie znanym zbiorem protokołów służących bezpiecznemu przesyłaniu przez sieć pakietów IP (ang. Internet Protocol) jest IPSec (ang. Internet Protocol Security). Zasada działania IPSec w zależności od wybranego trybu sprowadza się do szyfrowania całego pakietu IP i zaopatrzenia go w nowy nagłówek lub tylko do szyfrowania pola danych i przesyłania tak zmodyfikowanego pakietu przez sieć [15].

Innym rozwiązaniem, rozwijanym przez państwa członkowskie NATO, służącym do bezpiecznej wymiany połączeń głosowych i transmisji danych w relacjach "end-to-end" jest protokół SCIP (ang. Secure Communication Interoperability Protocol). Może być on wykorzystywany w sieciach telefonii komórkowej, ISDN (ang. Integrated Services Digital Network), PSTN (ang. Public Switched Telephone Network), VoIP (ang. voice over IP) [11]. Możliwe jest również wykorzystanie protokołu SCIP w radiostacjach szerokopasmowych, jednakże nie wszystkie tryby pracy z nim współpracują. Z tego też powodu za bezpieczeństwo komunikacji pakietowej IP z wykorzystaniem radiostacji odpowiada standard NINE (ang. NATO NII IP Network Encryption) oparty na Amerykańskim protokole HAIPE (ang. High Assurance Internet Protocol Encryptor). NINE wykorzystuje zestaw protokołów IPSec do zapewnienia szyfrowania i uwierzytelniania przesyłanych danych w sieci IP po uprzednim uwierzytelnieniu i uzgodnieniu klucza komunikujących się stron [2].

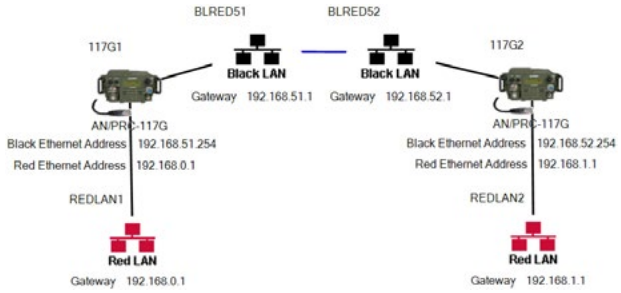
Układ pozostałej części artykułu zorganizowany został w następujący sposób. Rozdział 2 prezentuje architekturę proponowanego rozwiązania. Scenariusze badań wraz wynikami analizy opóźnienia, zmienności opóźnienia oraz przepływności łącza umieszczono w rozdziale 3. Opracowanie zakończone zostało wnioskami zamieszczonymi w sekcji 4.

2. KONCEPCJA WYKORZYSTANIA RADIOSTACJI WOJSKOWYCH

2.1. Radiostacja AN/PRC – 117G

Radiostacja AN/PRC-117G [9] jest nowoczesnym, szerokozakresowym i szerokopasmowym środkiem łączności współczesnego pola walki. Pracuje w zakresie częstotliwości od 30 do 1999,9950 MHz. Radiostacja ta jest urządzeniem klasy CCI (ang. Controlled Cryptographic Item) i posiada certyfikat Type-1, a więc najwyższy z możliwych, wydany przez Narodową Agencję Bezpieczeństwa (ang. National Security Agency, NSA) Stanów Zjednoczonych. Radiostacja zapewnia usługi transmisji fonii i danych w wąsko (do 25 kHz) i szerokopasmowych (do 5 MHz) kanałach radiowych. Dodatkową funkcją, jaką posiada ta radiostacja jest możliwość pracy jako certyfikowany przez NSA, szyfrator zapewniający bezpieczeństwo transmisji do klauzuli Top Secret włączenie.

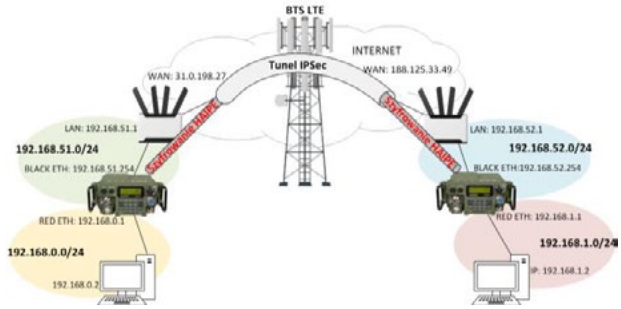
Programowanie radiostacji odbywa się w dedykowanej przez producenta (L3Harris Technologies, Inc.) aplikacji – Communications Planning Application (CPA). Graficzne zobrazowanie sposobu konfiguracji radiostacji pokazano na rys. 2. W rozpatrywanym scenariuszu radiostacja posiada dwa interfejsy sieciowe oznaczone jako Red i Black Ethernet. Sieć Red LAN reprezentuje nieszyfrowaną sieć lokalną. Elementy Black LAN reprezentują bezpieczne, szyfrowane połączenie pomiędzy radiostacjami.



Rys. 2. Schemat połączeń elementów w CPA

2.2. Architektura sieci

Rysunek 3 przedstawia sposób bezpiecznego połączenia dwóch sieci lokalnych (tutaj 192.168.0.0/24 oraz 192.168.1.0/24) z wykorzystaniem radiostacji AN/PRC-117G w roli szyfratora oraz infrastruktury sieci telefonii komórkowej LTE (ang. Long Term Evolution).



Rys. 3. Architektura sieci

Połączenie pomiędzy sieciami 192.168.51.0/24 oraz 192.168.52.0/24 zrealizowano przy wykorzystaniu modemów MikroTik Chateau 5G ax [10]. Są to dwuzakresowe access pointy z dodatkową obsługą standardów 3G, LTE i 5G. Modem pozwala tworzyć wirtualne sieci prywatne (ang. virtual private network, VPN) przy wykorzystaniu wielu protokołów, m.in.: PPTP, L2TP, OpenVPN, IPSec, SSTP, WireGuard. Urządzenia MikroTik pracują pod kontrolą systemu operacyjnego RouterOS (aktualnie w wersji 7.9). Przykładowe polecenia pozwalające na zestawienie tunelu VPN IPSec dla modemu umiejscowionego na rys. 3 z lewej strony przytoczone zostały na rys. 4. Należy tutaj jednoznacznie podkreślić, że tunel VPN IPSec ma za zadanie zapewnienie połączenia pomiędzy sieciami 192.168.51.0/24 oraz 192.168.52.0/24 i nie stanowi on podstawy bezpieczeństwa prezentowanego systemu.

3. WYNIKI

3.1. Scenariusz badań

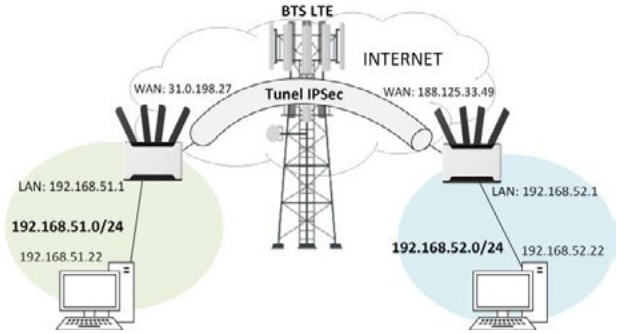
Celem badań jest określenie wpływu szyfrowania HAIPE na podstawowe parametry określające jakość usług sieciowych. Badania przeprowadzono w dwóch scenariuszach. Scenariusz pierwszy polegał na połączeniu dwóch sieci lokalnych poprzez sieć telefonii komórkowej. Schemat połączenia został przedstawiony na rys. 5. Drugi scenariusz stanowi rozbudowę scenariusza pierwszego poprzez umiejscowienie w strukturze sieci dwóch radiostacji AN/PRC-117. Schemat połączenia obrazuje rys. 3.

```
/ip ipsec profile add dh-group=modp2048 enc-algorithm=aes-128 name=ike1-site2

/ip ipsec proposal add enc-algorithms=aes-128-cbc name=ike1-site2 pfs-group=modp2048

/ip ipsec peer add address=188.125.33.49/32 name=ike1-site2 profile=ike1-site2

/ip ipsec identity add peer=ike1-site2 secret=haslowspoldzielone

/ip ipsec policy add src-address=192.168.51.0/24 src-port=any dst-address=192.168.52.0/24 dst-port=any tunnel=yes action=encrypt proposal=ike1-site2 peer=ike1-site2
```

Rys. 4. Przykładowa konfiguracja tunelu modemu

Do badania parametrów jakościowych wykorzystano dwa popularne narzędzia programowe:

- ping – do badania opóźnienia transmisji pakietów w sieci teleinformatycznej. Opóźnienie w tym wariancie jest rozumiane jako sumaryczny czas transmisji pakietów w obu kierunkach (ang. Round-Trip Time, RTT);
- iperf [3] – do badania przepływności w sieci teleinformatycznej oraz zmienności opóźnienia (ang. jitter).



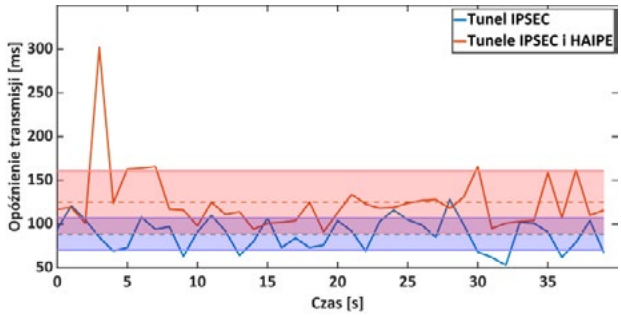
Rys. 5. Podstawowa architektura sieci – scenariusz 1

3.2. Opóźnienie transmisji

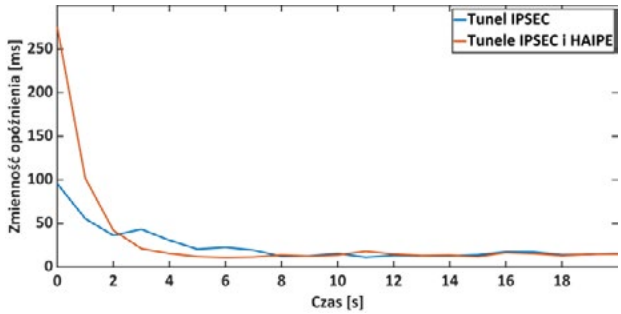
Na rys. 6 przedstawiono wyniki obrazujące zmianę wartości opóźnienia transmisji pakietów w czasie. Dla scenariusza pierwszego, który zakładał połączenie sieci lokalnych poprzez sieć telefonii komórkowej standardu LTE z jednoczesnym wykorzystaniem protokołu IPSec średnie opóźnienie transmisji wynosi 88,7 ms z odchyleniem standardowym wynoszącym 18,4 ms. Dodatkowe szyfrowanie HAIPE powoduje wzrost średniej wartości opóźnienia do poziomu 125,3 ms. Odchylenie standardowe w tym przypadku wynosi 35,5 ms.

3.1. Zmienność opóźnienia

Rysunek 7 przedstawia charakter zmian w czasie zmienności opóźnienia transmisji pakietów. Po ustabilizowaniu się badanej sieci wartości jittera dla obu analizowanych scenariuszy są zbliżone i oscylują wokół 14 ms.



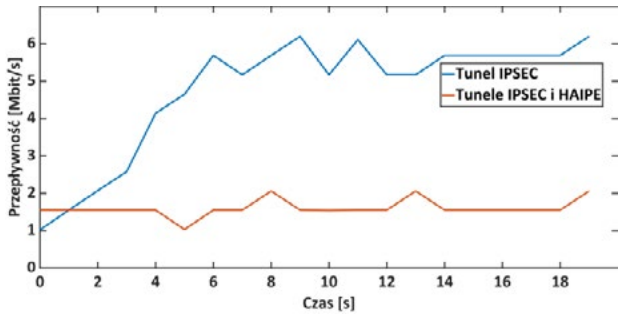
Rys. 6. Opóźnienie transmisji



Rys. 7. Zmienność opóźnienia transmisji

3.2. Przepływność łącza

Na rys. 8 przedstawiono zmianę przepływności w czasie w obu rozpatrywanych scenariuszach. Przepływność w sieci LTE z szyfrowaniem IPSec osiąga wartości na poziomie 5,6 Mbit/s. Dodatkowy tunel HAIPE znacząco obniża możliwość transferu danych pomiędzy użytkownikami sieci. W tym wariancie średnia przepływność tak zestawionego łącza kształtuje się na poziomie 1,6 Mbit/s.



Rys. 8. Szybkość transmisji danych

Narzędzie iperf dodatkowo zwraca informację o liczbie utraconych pakietów podczas transmisji. W trakcie prowadzonych badań nie odnotowano strat pakietów.

Dokumenty ITU-T G.1010 [5] oraz ITU-T Y.1541 [6] określają graniczne wartości metryk jakościowych.

Przykładowo dla usługi telefonii VoIP wartości opóźnienia transmisji, jittera, szybkości transmisji i strat pakietów wg ITU-T G.1010 wynoszą odpowiednio: 150ms, 1ms, 4-64 kbit/s i 3%. W przypadku wideorozmowy rekomendowana jest większa szybkość transmisji danych wynosząca 16-384 kbit/s i straty pakietów na poziomie 1%. Dla transmisji danych restrykcyjne są przede wszystkim wymagania dotyczące strat pakietów, które powinny być zerowe. Dokument ITU-T Y.1541 określa klasy QoS (ang. Quality of Service) wraz z przypisaniem do nich granicznych wartości parametrów jakościowych. Dla usług VoIP i wideorozmowy zalecana jest klasa 0 i 1. Dla tych klas wartości metryk jakościowych wynoszą 100 ms i 400 ms dla opóźnienia odpowiednio w klasie 0 i 1 oraz 50 ms dla zmienności opóźnienia (rozumianego tutaj jako IPDV [6] (ang. IP Packet Delay Variation)) w obu klasach i 10^{-3} dla strat pakietów.

Warto dodać, że wartości parametrów podane w ITU-T G.1010 odnoszą się do warstwy aplikacji, podczas gdy wartości zdefiniowane w ITU-T Y.1541 dotyczą warstwy sieci. Stąd też wynikają różnice pomiędzy wartościami niektórych z parametrów. Na poziomie aplikacji istnieją dodatkowe mechanizmy mogące zmienić charakterystyki strumienia pakietów [7]. Przykładowo na poziomie aplikacji graniczna wartość zmienności opóźnienia wynosi tylko 1 ms a na poziomie sieci 50 ms.

Parametry jakościowe, które zostały otrzymane podczas testów spełniają wymagania określone dla usług oferowanych w sieciach teleinformatycznych. Wykorzystując radiostacje jako nowoczesne i certyfikowane szyfratory klasy Type-1 oraz publiczną infrastrukturę sieci telefonii komórkowej można zapewnić bezpieczną wymianę danych pomiędzy dwoma niezależnymi sieciami teleinformatycznymi. Zaproponowany scenariusz pozwala na realizację usług teleinformatycznych z wysokim poziomem jakości i bezpieczeństwa.

4. WNIOSKI

Artykuł przedstawia koncepcję i realizację bezpiecznego systemu łączności specjalnej. Zaprezentowano rozwiązanie, które wykorzystuje istniejącą infrastrukturę sieci telefonii komórkowej oraz radiostacje wojskowe jako wysoce bezpieczne szyfratory, certyfikowane przez upoważnione do tego służby. Przedstawione w artykule wyniki pokazują, że zaproponowane rozwiązanie umożliwia wymianę danych i realizację usług pomiędzy węzłami sieci.

W ramach dalszych prac przewiduje się realizację scenariusza, który wykorzystywał będzie sieć 5G. Planuje się również zbudowanie systemu bezpiecznej wymiany danych w oparciu o sieć łączności satelitarnej.

LITERATURA

[1] Agencja Bezpieczeństwa Wewnętrznego. *Bezpieczeństwo Teleinformatyczne.* 2023. https://bip.abw.gov.pl/bip/informacje-niejawne-1/bezpieczenstwo-teleinf/154,BEZPIECZENSTWO-TELEINFORMATYCZNE.html (dostęp: 18.05.2023).

[2] Białas Robert, Grzonkowski Marcin i Wicik Robert. „Cryptographic Protection for Military Radio Communications." *INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS.* Polish Academy of Sciences Committee of Electronics and Telecommunications, 2020. 687-693.

[3] Energy Sciences Network, Lawrence Berkeley National Laboratory. *iPerf - The ultimate speed test tool for TCP, UDP and SCTP.* 8 Czerwiec 2016. https://iperf.fr/ (dostęp: 17.05.2023).

[4] Grzonkowski Marcin, Wicik Robert i Białas Robert. „Cryptographic Protection for Military Radio Communications." *INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS.* Polish Academy of Sciences Committee of Electronics and Telecommunications, 2020. 687-693.

[5] International Telecommunication Union. *End-user multimedia QoS categories.* ITU-T Recommendation G.1010, International Telecommunication Union, 2002.

[6] International Telecommunication Union. *Network performance objectives for IP-based services.* ITU-T Recommendation Y.1541, International Telecommunication Union, 2012.

[7] Janowski Robert. „Ocena jakości usług telefonii pakietowej." *Zeszyty Naukowe Warszawskiej Wyższej Szkoły Informatyki*, 2013: 7-24.

[8] Kaleta Sebastian. „Bezpieczeństwo informacji niejawnych w systemach i sieciach teleinformatycznych Sił Zbrojnych Rzeczypospolitej Polskiej." *Wiedza Obronna*, 2014: 16-33.

[9] L3Harris Technologies, Inc. „L3HARRIS FALCON III® AN/PRC-117G(V)1(C) Multiband Networking Manpack Radio." *AN/PRC-117G(V)1(C) Multiband Networking Manpack Radio.* Październik 2019. (dostęp: 17.05.2023).

[10] MikroTik. *Chateau 5G ax.* 2023. https://mikrotik.com/product/chateau_5g_ax#fndtn-specifications (dostęp: 17.05.2023).

[11] Powiertowski Wojciech, Szuster Jan i Kaczkowski Maciej. „Moduł kryptograficzny w zastosowaniach w telefonie AC-16SCIP oraz urządzeniu IPSec." *Konferencja Zastosowań Matematyki.* Zakopane, 2011.

[12] „Rozporządzenie Prezesa Rady Ministrów z dnia 20 lipca 2011 r. w sprawie podstawowych wymagań bezpieczeństwa teleinformatycznego." Dz.U. 2011 nr 159 poz. 948, 2011.

[13] Służba Kontrwywiadu Wojskowego. *Bezpieczeństwo teleinformatyczne.* 2023. https://bip.skw.gov.pl/skw/bezpieczenstwo-teleinfo/5107,Bezpieczenstwo-teleinformatyczne.html (dostęp: 18.05.2023).

[14] „Ustawa z dnia 5 sierpnia 2010 r. o ochronie informacji niejawnych." Dz.U. 2010 nr 182 poz. 1228, 2010.

[15] Wikipedia. *IPsec.* 11 Kwiecień 2023. https://pl.wikipedia.org/wiki/IPsec#Linki_zewnętrzne (dostęp: 17.05.2023).

INFRASTRUKTURY KOMUNIKACJI KWANTOWEJ – SYNERGIE ORAZ INTEGRACJA Z SYSTEMAMI TRANSMISJI WZORCOWYCH SYGNAŁÓW CZASU I CZĘSTOTLIWOŚCI

QUANTUM COMMUNICATION INFRASTRUCTURES – SNERGIES AND INTEGRATION WITH REFERENCE TIME AND FREQUENCY TRANSMISSION SYSTEMS

Piotr Rydlichowski[1]; Adrian Kliks[2]

[1] Poznańskie Centrum Superkomputerowo-Sieciowe, Poznań, Piotr.rydlichowski@man.poznan.pl
[2] Politechnika Poznańska, Poznań, adrian.kliks@put.poznan.pl

DOI: 10.15199/59.2023.4.26

Streszczenie: **Technologie obliczeń oraz komunikacji kwantowej spotkały się w ostatnich latach z dużym zainteresowaniem. Postępy w technologiach fotonicznych pozwalają na wdrażanie stanowisk testowych komunikacji kwantowej, w których urządzenie oparte na źródłach i detektorach pojedynczych fotonów i splątanie kwantowe są wykorzystywane do przesyłania bitów kwantowych i oferują nowe możliwości zastosowania i usługi, również w połączeniu z istniejącymi systemami i usługami optycznej transmisji danych. W pracy przedstawiono stan bieżący integracji infrastruktury komunikacji kwantowej z systemami transmisji sygnałów wzorcowych.**

Abstract: **Quantum Computing and Quantum Communication technologies received significant interest over recent years. Advancements in photonic technologies allow for the implementation of quantum communication testbeds where single photon devices and quantum entanglement is used to transmit quantum bits and offer new possible use cases and services also with connection with existing optical data transmission systems and services. This paper analyzes the integration of the quantum communication systems with the reference signal systems.**

Słowa kluczowe: **QKD, komunikacja kwantowa, obliczenia kwantowe, kwantowa dystrybucja klucza, sygnały czasu i częstotliwości**

Keywords: **QKD, quantum communication, quantum computing, quantum key distribution, time and frequency signals**

1. WSTĘP

W ostatnich latach coraz więcej zasobów oraz czasu poświęca się tematyce związanej z sieciami komunikacji kwantowej oraz koncepcjom Internetu kwantowego – zagadnieniami tymi zajmuje się między innymi grupa robocza IETF - *Quantum Internet Research Group* [1]. Pierwszą użyteczną potencjalną usługą wykorzystującą schemat komunikacji kwantowej jest technologia Quantum Key Distribution (QKD) [2], [3]. Systemy QKD pozwalają na generowanie i dystrybucję z wykorzystaniem zasad komunikacji kwantowej kluczy, które mogą być wykorzystywane w różnych aplikacjach i usługach (np. szyfrowanie czy uwierzytelnianie). Istniejące systemy QKD wykorzystują albo ciemne włókna, albo mechanizm wspólnej propagacji kanału kwantowego w ramach istniejącego widma systemu transmisji optycznej. Technologia i infrastruktura QKD to pierwszy krok w kierunku ogólnych systemów komunikacji kwantowej. Na całym świecie powstało szereg infrastruktur testowych i projektów QKD, a systemy QKD zostały wdrożone w operacyjnych środowiskach sieciowych. Aktualne inicjatywy takie jak EURO-QCI (*European Quantum Communication Infrastructure*) [4] ukierunkują badania i rozwój w kierunku dużych krajowych i międzynarodowych infrastruktur testowych komunikacji kwantowej (z wykorzystaniem zarówno naziemnych systemów optycznych, jak i satelitarnego schematu komunikacji kwantowej). Powyższe systemy i infrastruktury pozwolą na dalszy rozwój technologii komunikacji kwantowej, standaryzację, certyfikację i wdrażanie schematów użycia. Uważa się, że infrastruktury testowe QCI będą wykorzystywać rozwiązania technologii mieszanej, wykorzystujące specjalnie do tego celu przeznaczone ciemne światłowody lub istniejące współdzielone widmo optycznych systemów transmisji danych. Systemy komunikacji kwantowej z ciemnymi włóknami oferują lepszą wydajność, taką jak zasięg i kwantową bitową stopę błędów dla usług komunikacji kwantowej. Ze względu na swoją naturę niektóre elementy i procedury sieciowe nie mogą być zaimplementowane w czysto kwantowej domenie komunikacyjnej i w związku z tym hybrydowe rozwiązania klasyczno-kwantowe są planowane do zastosowania w sieciach komunikacji kwantowej.

Infrastruktury komunikacji kwantowej oferują jednocześnie znaczące nowe możliwości dla istniejących i przyszłych systemów transmisyjnych z dystrybucją wzorcowych sygnałów czasu i częstotliwości. Z jednej strony infrastruktury te mogą korzystać ze wspólnych zasobów, a z drugiej strony systemy komunikacji kwantowej będą wymagały różnych poziomów dokładności sygnałów synchronizacji czasu i częstotliwości, aby prawidłowo działać i osiągać optymalną wydajność. Kanały komunikacji kwantowej oraz wzorcowe sygnały czasu i częstotliwości można traktować jako szczególny rodzaj nowych sygnałów wykorzystywanych i przesyłanych w klasycznych optycznych systemach transmisji danych. Obie usługi mogą w znacznym stopniu skorzystać ze wspólnych, współdzielonych zasobów infrastruktury oraz rozszerzyć możliwy zasięg i dostępność obu usług. Potencjalnie możliwe są nowe rozwiązania, ponieważ systemy i kanały komunikacji kwantowej dostarczają dokładniejsze dane w czasie

rzeczywistym i monitorowanie dotyczące ścieżki optycznej i segmentu światłowodu. Ponadto systemy komunikacji kwantowej mogą zwiększyć bezpieczeństwo dystrybucji i transmisji wzorcowych sygnałów czasu i częstotliwości. Innym aspektem jest integracja sieci QCI na dużą skalę oraz sieci wzorcowych sygnałów czasu i częstotliwości, współistnienie i współdzielenie zasobów, zwłaszcza w świetle wymagań ciemnych włókien.

W artykule przedstawiono możliwe scenariusze integracji infrastruktury QCI (kanały komunikacji kwantowej) z istniejącymi referencyjnymi systemami transmisyjnymi dystrybucji czasu i częstotliwości oraz przyszłymi rozwiązaniami, takimi jak referencyjna nośna optyczna. Przedstawiono scenariusze integracji systemów – zarówno z wykorzystaniem ciemnego światłowodu, jak i wariantu współdzielonego widma optycznego. Omówiono perspektywę przyszłości QKD (CV-QKD, MDI-QKD, TF-QKD) oraz ogólną komunikację kwantową.

2. Podstawy QKD

Głównymi założeniami schematu komunikacji kwantowej jest praktyczna implementacja różnych technik transmisji wykorzystujących kodowanie danych za pomocą stanów kwantowych. Te elementy tworzą bazę dla transmisji kwantowej, sieci kwantowych i Internetu kwantowego. Pierwszą użyteczną praktyczną implementacją komunikacji kwantowej jest technologia Quantum Key Distribution [2], [3]. Prezentuje ona, w jaki sposób fotony mogą być przesyłane, odbierane i wykorzystywane do osiągnięcia wartości dodanej w postaci bezpieczeństwa. Systemy QKD mogą wykorzystywać różne założenia bazowe, w wyniku czego dostępne są różne implementacje protokołów, takie jak BB84, BB92, czy COW [5]. Mogą one osiągnąć różną wydajność i poziom złożoności. Następna generacja systemów QKD będzie wykorzystywać bardziej zaawansowany schemat wykorzystujący splątanie kwantowe i teleportację. Obecnie możemy wyróżnić dwa główne typy systemów QKD: DV-QKD i CV-QKD.

2.1. DV-QKD

Metoda *discrete-variable QKD* (DV-QKD) [6]wymaga detektorów optycznych do identyfikacji pojedynczych fotonów (pomiary dyskretne: tak/nie). Aby wdrożyć takie rozwiązania na duże odległości, potrzebne są tzw. regeneratory kwantowe, których technologia wymaga jeszcze znacznego rozwoju. System *discrete-variable* opiera się na wykrywaniu pojedynczych fotonów , a także może wykorzystywać splątanie kwantowe i teleportację informacji. Wynikiem transmisji i analizy jest wartość dyskretna – czy foton został wykryty, czy nie. Technika DV-QKD jest łatwiejsza do wdrożenia, a pierwsze systemy QKD zostały już zaimplementowane na szeregu platformach sprzętowych. Ponadto źródła i detektory pojedynczych fotonów mogą być wykorzystywane w różnych innych zastosowaniach. Wadą tej techniki jest to, że urządzenia z pojedynczym fotonem wymagają ścisłych warunków środowiskowych i nie są kompatybilne z najnowszymi schematami transmisji optycznej. Powoduje to, że w większości przypadków musimy użyć ciemnych włókien optycznych przeznaczonych wyłącznie dla tych systemów transmisji.

2.2. CV-QKD

Inna metoda nosi nazwę *ciągłej zmiennej QKD* (CV-QKD) [7]. Systemy CV-QKD oparte są na detekcji homodynowej i wykonują kwadraturowe pomiary pola elektrycznego padającego światła. Wynikiem tej transmisji i pomiaru jest rzutowanie fazy i amplitudy pola elektrycznego światła na osie kwadraturowe i takie podejście jest typowe dla każdego spójnego systemu transmisyjnego. Uzyskane wyniki wskazują, że są to wartości ciągłe w czasie. W takim spójnym podejściu bezpieczeństwo transmisji opiera się na kwantowych obliczeniach statystycznych, a nie na mechanizmach splątania i teleportacji. Większość obecnie dostępnych urządzeń QKD bazuje na takiej implementacji CV-QKD. Główną zaletą tych systemów jest wyższa wydajność oraz kluczowa możliwość integracji z istniejącymi systemami optycznej transmisji danych. Wadą takiego podejścia jest to, że system jest bardziej złożony i wymaga mniejszego budżetu mocy optycznej dla segmentów optycznych. Z punktu widzenia implementacji różnicę między DV-QKD a CV-QKD można postrzegać jako różnicę między detektorem pojedynczego fotonu a detektorem homodynowym. Ta analogia jest taka sama, jak w przypadku obecnych optycznych systemów transmisji danych.

3. Transmisja wzorcowych sygnałów czasu i częstotliwości

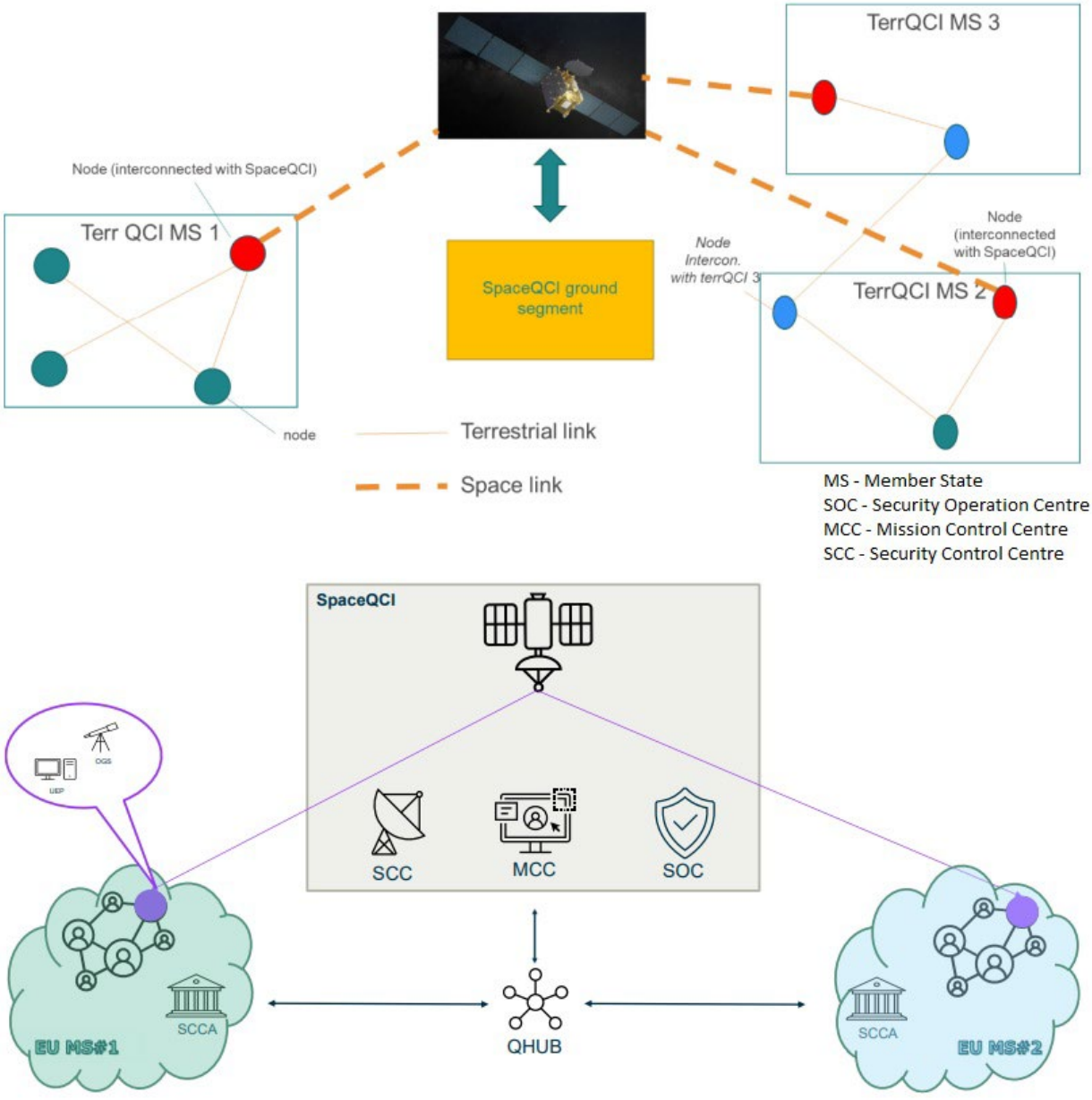
W ostatnich latach opracowano wiele różnych technik dystrybucji i wysyłania wzorcowych sygnałów czasu i częstotliwości. Są one zwykle oparte na źródle, takim jak zegary atomowe lub tzw. zegary optyczne. Główną ideą jest takie rozprowadzenie tych sygnałów, aby były użyteczne i możliwe do porównania w różnych lokalizacjach bez znaczącej utraty krótko- i długoterminowej stabilności. Głównym problemem jest zaprojektowanie systemu transmisyjnego w sposób umożliwiający monitorowanie i kompensację wpływu różnych czynników, takich jak np. opóźnienie, temperatura, dyspersja, które zmieniają się podczas transmisji sygnału optycznego. Generuje to dwa podejścia do takich systemów transmisji: mogą one wykorzystywać dedykowane ciemne światłowody lub współdzielić widmo optycznego systemu transmisji danych. System oparty na podejściu światłowodowym charakteryzuje się najlepszą wydajnością i dokładnością. Wzorcowe sygnały czasu i częstotliwości mogą być wykorzystywane w wielu zastosowaniach, takich jak na przykład geodezja i synchronizacja krytycznych systemów transmisyjnych.

4. QKD i integracja z systemami transmisji wzorcowych sygnałów czasu i cętotliwości

Z przedstawionych powyżej uwag można wywnioskować, że z praktycznego punktu widzenia te dwie techniki stwarzają podobne wyzwania: są to nowe rodzaje specjalnych sygnałów i mogą być transmitowane ciemnym światłowodem lub w widmie istniejącego optycznego systemu transmisji danych. Kompromis dotyczy wydajności i zasięgu. Szereg projektów i inicjatyw, zwłaszcza w Unii Europejskiej, proponuje tworzenie specjalistycznych sieci i systemów dla tych dwóch rodzajów specjalnych sygnałów optycznych. Sieci te będą miały kluczowe znaczenie

dla dalszego rozwoju tych technologii i wdrażania bardziej praktycznych zastosowań i schematów użycia. Ponieważ najlepszą wydajność tych systemów można osiągnąć za pomocą dedykowanego światłowodu, jest to preferowany sposób i można dodatkowo wykorzystać wspólne zasoby. Należy podkreślić, że systemy QKD wymagają również synchronizacji pomiędzy węzłami końcowymi. Z drugiej strony systemy QKD ze względu na swój charakter dokładnie monitorują stan toru optycznego, a zmiany są odzwierciedlane w wydajności i monitorowane. Daje to możliwość wykorzystania tych danych do dalszej kompensacji potencjalnych zmian w systemach transmisji wzorcowych sygnałów czasu i częstotliwości, biorąc pod uwagę, że używamy tej samej ścieżki optycznej i segmentów. Proponowana technika polega na multipleksowaniu i demultipleksowaniu w każdym węźle sieci optyczne takiego systemu musi być odpowiednio skonfigurowane, ponieważ na systemy QKD nie może mieć wpływu klasyczne optyczne wzmocnienie sygnałów (wzmacniacze typu EDFA). Systemy CV-QKD mają tę zaletę, że ze względu na swój charakter można je precyzyjnie dostroić do określonej długości fali. Inicjatywa EuroQCI – European Quantum Communication Infrastructure jest w trakcie przygotowywania szczegółowej architektury systemów i wykorzystania również połączeń transgranicznych w postaci naziemnych łączy optycznych lub satelitarnej komunikacji kwantowej w wolnej przestrzeni. Na rysunku 1 przedstawiono proponowaną przez Komisję Europejską konfigurację łączy QCI-QKD oraz poszczególne elementy systemu [8].

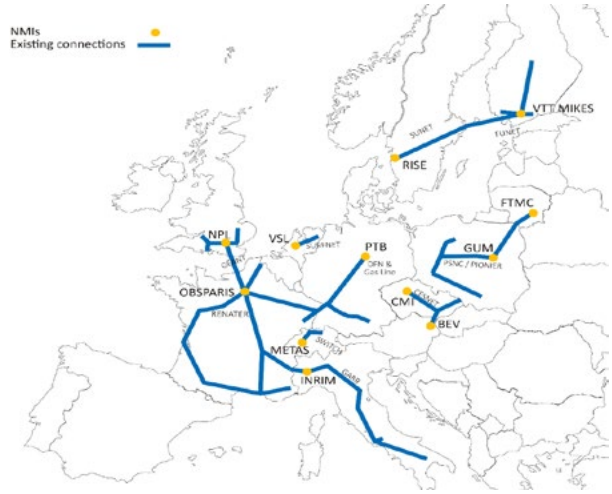
W odniesieniu do systemów transmisji wzorcowych sygnałów czasu i częstotliwości w Europie istnieją już lo-



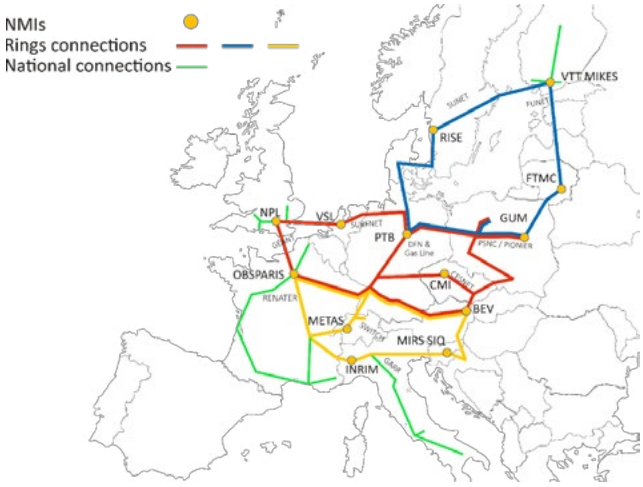
tych dwóch rodzajów systemów transmisyjnych. Widmo

Rys. 1. Proponowana konfiguracja EuroQCI [8]

kalne połączenia lub w obrębie danego kraju członkowskiego natomiast brakuje systemowego podejścia oraz infrastruktury umożliwiającej transmisję oraz porównywanie tych sygnałów w skali europejskiej, szczególnie pomiędzy wszystkimi krajowymi instytutami metrologicznymi. Głównym wyzwaniem jest brak infrastruktury światłowodowej umożliwiającej długodystansową, bezpośrednią transmisję sygnałów pomiędzy partnerami. Rysunek 2 [9] przedstawia wymienione powyżej istniejące połączenia. Rysunek 3 [9] przedstawia natomiast proponowaną konfigurację w skali europejskiej.



Rys. 2. Istniejące połączenia dla transmisji wzorcowych sygnałów czasu i częstotliwości [9]



Rys. 3. Proponowane połączenia dla transmisji wzorcowych sygnałów czasu i częstotliwości [9]

5. Podsumowanie

Niniejszy artykuł przedstawia problematykę rozwoju i integracji nowego typu usług i sygnałów jakimi są systemu komunikacji kwantowej oraz transmisji wzorcowych sygnałów czasu i częstotliwości. Dla dalszego rozwoju i integracji wymagają dedykowanej operacyjnej oraz laboratoryjnej infrastruktury telekomunikacyjnej. W celu zoptymalizowania zasobów zasadnym jest integracja tych dwóch typów usług oraz sygnałów oraz traktowanie ich jako specjalnego rodzaju sygnałów przesyłanych w systemach telekomunikacyjnych.

LITERATURA

[1] IETF, Qirg -- Quantum Internet RG; https://www.irtf.org/mailman/listinfo/qirg (dostep 8 maja 2023)

[2] P. Sharma, A. Agrawal, V. Bhatia, S. Prakash and A. K. Mishra, "Quantum Key Distribution Secured Optical Networks: A Survey," in IEEE Open Journal of the Communications Society, vol. 2, pp. 2049-2083, 2021. doi: 10.1109/OJCOMS.2021.3106659

[3] Y. Cao, Y. Zhao, Q. Wang, J. Zhang, S. X. Ng and L. Hanzo, "The Evolution of Quantum Key Distribution Networks: On the Road to the Qinternet," in IEEE Communications Surveys & Tutorials, vol. 24, no. 2, pp. 839-894, Secondquarter 2022. doi: 10.1109/COMST.2022.3144219

[4] Komisja Europejska, EuroQCI, https://digital-strategy.ec.europa.eu/pl/policies/european-quantum-communication-infrastructure-euroqci (dostep 8 maja 2023)

[5] S. R. Hasan, M. Z. Chowdhury, M. Saiam and Y. M. Jang, "Quantum Communication Systems: Vision, Protocols, Applications, and Challenges," in IEEE Access, vol. 11, pp. 15855-15877, 2023, doi: 10.1109/ACCESS.2023.3244395.

[6] Ivan B. Djordjevic, Discrete Variable (DV) QKD, In: Physical-Layer Security and Quantum Key Distribution, pp. 267-322, 2019, https://link.springer.com/chapter/10.1007/978-3-030-27565-5_7

[7] Fabian Laudenbach, Christopher Pacher, Chi-Hang Fred Fung, Andreas Poppe, Momtchil Peev, Bernhard Schrenk, Michael Hentschel, Philip Walther, Hannes Hübel, Continuous-Variable Quantum Key Distribution with Gaussian Modulation – The Theory of Practical Implementations, 2018,https://arxiv.org/abs/1703.09278

[8] Euro QCI Concept of Operations (ConOps) https://digital-strategy.ec.europa.eu/en/euroqci-conops-concept-operations

[9] Wojbor Bogacki PTTI 2023 - Long Beach, CA, January 23-26, 2023 CLONETS-DS – Clock Network Services-Design study Strategy and innovation for clock services over optical-fibre networks

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

DEKOMPOZYCJA IMPULSÓW RADAROWYCH ZA POMOCĄ TRIANGULACJI ZER SPEKTROGRAMU

RADAR PULSE DECOMPOSITION USING SPECTROGRAM ZERO TRIANGULATION

Karol Abratkiewicz

Instytut Systemów Elektronicznych,
Politechnika Warszawska,
00-665 Warszawa
karol.abratkiewicz@pw.edu.pl

DOI: 10.15199/59.2023.4.27

Streszczenie: W niniejszej pracy przedstawiono wyniki analizy i dekompozycji impulsów radarowych za pomocą metody triangulacji zer spektrogramu w dziedzinie czas-częstotliwość. Sygnał niestacjonarny zaburzony szumem białym charakteryzuje się ściśle określonym rozkładem zer na płaszczyźnie czas-częstotliwość, jeśli do przekształcenia wykorzysta się okno Gaussa oraz krótkoczasową transformację Fouriera (ang. short-time Fourier transform, STFT). Wówczas odległości pomiędzy sąsiadującymi zerami spektrogramu (rozkładu energii STFT) wykazują się inną dystrybucją dla szumu i inną w otoczeniu sygnału. Dzięki tej właściwości możliwe było pokazanie w artykule procesu ekstrakcji impulsów pochodzących z nadajników radiolokacyjnych.

Abstract: This paper presents the results of the analysis and decomposition of radar pulses using the method of triangulation of spectrogram zeros in the time-frequency domain. A non-stationary signal disturbed by white noise is characterized by a strictly defined distribution of zeros on the time-frequency plane if the Gaussian window and the short-time Fourier transform (STFT) are used for the analysis. Then the distances between adjacent zeros of the spectrogram (STFT energy distribution) are characterized by a different distribution for noise and in the signal vicinity. Thanks to this property, the article shows the process of extracting pulses from radar transmitters.

Słowa kluczowe: analiza sygnałów w dziedzinie czas-częstotliwość, radar, radiolokacja, STFT

Keywords: radar, radiolocation, STFT, time-frequency analysis

1. WSTĘP

Analiza sygnałów niestacjonarnych w dziedzinie czas-częstotliwość jest zagadnieniem, które rozwija się intensywnie, co można zaobserwować szczególnie w ostatnich latach [1, 2, 3, 4]. Wynika to między innymi z rosnącej wydajności obliczeniowej komputerów i współpracujących podzespołów (np. kart graficznych wykorzystywanych do obliczeń równoległych), dzięki której możliwa jest wydajna implementacja zaawansowanych algorytmów cyfrowego przetwarzania sygnałów w czasie rzeczywistym [5]. Szybka i efektywna realizacja złożonych metod analizy sygnałów ma obecnie wyjątkowo duże znaczenie w rozwiązaniach specjalnych i profesjonalnych, w których niezawodność i jak najniższa latencja to zwykle kluczowe wymagania.

Czasowo-częstotliwościowe metody przetwarzania sygnałów są przydatnym narzędziem służącym do analizy i dekompozycji sygnałów radiowych, w tym radarowych [6, 5]. Techniki, takie jak np. STFT, transformacja falkowa czy przekształcenie Wigner Ville'a, pozwalają na wydobycie zależności czasowych i częstotliwościowych jednocześnie. Dzięki transformacji sygnału do płaszczyzny czas-częstotliwość, można dokonać rozróżnienia komponentów, które zajmują to samo pasmo lub ten sam interwał czasowy, jednak stanowią osobne składowe sygnału. W literaturze można znaleźć przykłady takich algorytmów dekompozycji sygnałów, dzięki którym sygnał radarowy został rozłożony na składowe za pomocą technik czasowo-częstotliwościowych [5]. W niniejszej pracy wykorzystano podejście oparte o triangulację zer spektrogramu (dystrybucji energii sygnału na płaszczyźnie czas-częstotliwość). Ten wydajny algorytm został zaproponowany w [4] i wykorzystany do dekompozycji sygnałów echolokacyjnych generowanych przez nietoperze. W rezultacie otrzymano odseparowane (czasowo i częstotliwościowo) i odfiltrowane składowe sygnału wysyłanego przez te zwierzęta. Niniejszy artykuł przedstawia adaptację tej techniki na potrzeby filtracji impulsów radiolokacyjnych rejestrowanych przez odbiorniki sygnałów radiowych i radary aktywne i pasywne. W przypadku impulsów radarowych wysyłanych przez systemy dalekiego zasięgu (w tym systemów pozahoryzontalnych) moc sygnału zwykle sięga setek kW lub więcej. Detekcja impulsu pochodzącego z takiego radaru możliwa jest nawet z odległości setek lub tysięcy kilometrów (w przypadku propagacji jonosferycznej lub troposferycznej). Problemem może jednak być niska jakość odebranego sygnału. O ile w procesie detekcji obiektów ruchomych w klasycznych radarach przeszkodą jest przede wszystkim szum (termiczny lub pochodzący z zakłóceń), w odbiornikach rozpoznania elektronicznego na jakość impulsu ma wpływ nie tylko szum, ale także zjawisko propagacji wielodrogowej. Kwestia ekstrakcji impulsu zakłóconego przez szum i zniekształconego przez propagację wielodrogową została przeanalizowana i rozwiązana w tym artykule za pomocą techniki triangulacji zer spektrogramu. W konsekwencji dokonano separacji czasowej i częstotliwościowej impulsu radiolokacyjnego.

Dalsza część artykułu ma następującą strukturę: w

Sekcji 2 przedstawiono podstawy teoretyczne i opis badanego algorytmu. Następnie w Sekcji 3 zaprezentowano wyniki przetwarzania sygnałów pochodzących z pomiarów. Artykuł jest zakończony podsumowaniem i wskazaniem potencjalnych możliwości dalszego rozwoju algorytmu.

2. DEKOMPOZYCJA SYGNAŁU Z WYKORZYSTANIEM TRIANGULACJI ZER SPEKTROGRAMU

Jak wspomniano w Sekcji 1, technika ekstrakcji komponentów ze spektrogramu została zaproponowana przez Flandrina w [4]. Zakłada ona, że dla sygnału $x(t)$ oraz okna analizy $h(t)$ STFT zdefiniowana jest następująco:

$$F_x^h(t,\omega) = \int_{\mathbb{R}} x(\tau)h^*(\tau - t)e^{-j\omega\tau}\mathrm{d}\tau, \tag{1}$$

gdzie $j = \sqrt{-1}$, s^* oznacza sprzężenie s, a $\mathbb{R}$ to zbiór liczb rzeczywistych. Spektrogram, czyli rozkład energii sygnału na płaszczyźnie czas-częstotliwość, zdefiniowany jest jako kwadrat modułu STFT:

$$E_x^h(t,\omega) = \left|F_x^h(t,\omega)\right|^2. \tag{2}$$

STFT jest transformacją odwracalną, dzięki czemu dwuwymiarowa dystrybucja (1) może być zrekonstruowana do postaci sygnału w dziedzinie czasu za pomocą odwrotnej transformacji:

$$x(t - t_0) = \frac{1}{2\pi \cdot h^*(t_0)} \int_{\mathbb{R}} F_x^h(t,\omega)e^{j\omega(t-t_0)}\mathrm{d}\omega, \tag{3}$$

gdzie $h(t)$ jest niezerowe i ciągłe dla każdego t_0.

STFT (1) jest dystrybucją sygnału na płaszczyźnie czas-częstotliwość i może być analizowana jako rozkład funkcji zespolonej $z = \omega + jt$. Zakładając, że wykorzystywanym oknem analizy jest funkcja Gaussa o jednostkowej energii dana jako:

$$h(t) = \pi^{-1/4}e^{-\frac{t^2}{2}}, \tag{4}$$

STFT (1) może być zapisana jako:

$$F_x^h(t,\omega) = e^{-\frac{|z|^2}{4}}\mathcal{F}_x(z), \tag{5}$$

gdzie

$$\mathcal{F}_x(z) = \int_{\mathbb{R}} A(z,u)x(u)\mathrm{d}u, \tag{6}$$

to transformacja Bargmanna [7] o jądrze przekształcenia:

$$A(z,u) = \pi^{-\frac{1}{4}}e^{-\frac{u}{2}-juz+\frac{z^2}{4}}. \tag{7}$$

Transformacja Bargmanna (6) umożliwia faktoryzację Weierstrassa-Hadamarda daną jako:

$$\mathcal{F}_x(z) \propto \prod_{n=1}^{\infty}\left(1 - \frac{z}{z_n}\right)e^{\frac{z}{z_n}+\frac{1}{2}(\frac{z}{z_n})^2}, \tag{8}$$

gdzie $z_n = \omega_n + jt_n$ oznacza zera transformacji Bargmanna (6). Z (5) oraz (2) wynika, że zera obu funkcji są tożsame, a spektrogram może być zdefiniowany przez ich rozkład. Wówczas rozkład zer spektrogramu określa dystynktywne właściwości dystrybucji czasowo-częstotliwościowej sygnału. Ta właściwość została wykorzystana w [4] do dekompozycji sygnału wielokomponentowego.

Flandrin zauważył, że w przypadku szumu białego, odległość euklidesowa między sąsiadującymi zerami jest ograniczona od góry $L_{\max} \sim 2.16$ z prawdopodobieństwem przekroczenia wartości 2 wynoszącym 10^{-3}. W praktyce dodatkowo stosuje się ograniczenie od dołu $L_{\min} < |e_{m,n}|$, dla $L_{\min} \sim 0.2$. W myśl algorytmu zaproponowanego w [4] sąsiadujące zera spektrogramu są łączone za pomocą triangulacji Delaunaya, dając trójkąty o bokach z zakresu $L_{\min} < |e_{m,n}| < L_{\max}$ dla szumu białego, oraz spoza tego zakresu, w przypadku wystąpienia sygnału na płaszczyźnie czas-częstotliwość. Na podstawie przyjętych założeń dotyczących długości boku trójkąta wybierane są tylko te trójkąty, których co najmniej jeden bok ma długość spoza zakresu definiującego szum. Sąsiadujące trójkąty o bokach o długości wykraczającej poza zakres długości dla szumu, są łączone. Zbiór takich trójkątów definiuje maskę binarną (1 – wykryty komponent, 0 – szum). Stosując maskę binarną na pierwotnym rozkładzie STFT dokonuje się selekcji komponentów, które następnie są rekonstruowane za pomocą odwrotnej transformacji STFT danej zależnością (3).

3. WYNIKI ANALIZY

W niniejszej sekcji zaprezentowano wyniki analizy sygnału pochodzącego z nadajnika radiolokacyjnego. Niekooperujący radar zlokalizowany ok. 2 km od stacji odbiorczej emitował impulsy w celu detekcji obiektów w otaczającej przestrzeni. W tym samym czasie rejestrator dokonał akwizycji sygnału, który następnie odfiltrowano, zdecymowano i poddano dalszemu przetwarzaniu. Ze względu na niejawny charakter nadajnika, parametry impulsu nie zostały przedstawione i posłużono się unormowanymi wielkościami. Nie wpływa to na badany proces ekstrakcji sygnatury radaru ani przedstawione wyniki.

Na Rysunku 1 przedstawiono przebieg wyselekcjowanego impulsu w dziedzinie czasu. Zaprezentowano impuls w postaci części rzeczywistej ($\Re$) oraz urojonej ($\Im$). Widoczny na wykresie impuls jest wyraźny, jednak ze względu na obecność szumu i niejednorodość obwiedni, nie jest możliwe wyodrębnienie go z zaprezentowanego przebiegu. Ponieważ założeniem analizy była filtracja sygnału z liniową modulacją częstotliwości, możliwe było dobranie okna $h(t)$ w taki sposób, aby uzyskać optymalną roozdzielczość dystrybucji sygnału na płaszczyźnie czas-częstotliwość, mimo braku informacji na temat parametrów sygnału. Przyjęto, że sygnał z liniową modulacją częstotliwości zdefiniowany jest przez model:

$$y(t) = A(t)e^{j\left(\phi_0+\omega_0 t+\frac{\alpha t^2}{2}\right)}, \tag{9}$$

gdzie $A(t)$ to amplituda, ϕ_0 to faza początkowa, $\omega_0 = 2\pi f_0$ to pulsacja (przy czym f_0 to częstotliwość nośna sygnału), a α to współczynnik modulacji częstotliwości. Dla takiego sygnału i gaussowskiego okna (4) optymalny współczynnik szerokości okna dany jest następująco [8]:

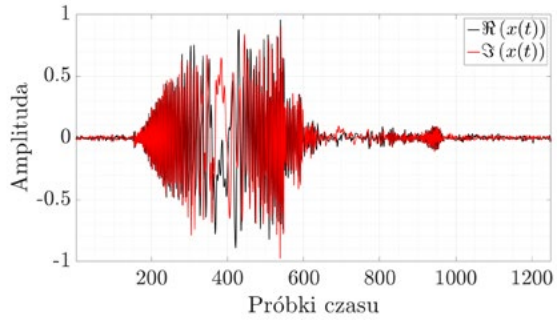
$$\sigma = \frac{1}{\sqrt{|\alpha|}}. \tag{10}$$

Współczynnik α nie był znany, jednak możliwe było określenie optymalnej wartości σ wykorzystując podejście zaproponowane w [9], w którym znaleziono zależność między α i σ. W tym celu posłużono się entropią Rényiego – klasyczną miarą koncentracji (zawartości informacji) rozkładu sygnału na płaszczyźnie czas-częstotliwość [10]. Entropia Rényiego rzędu γ dowolnej dystrybucji sygnału na

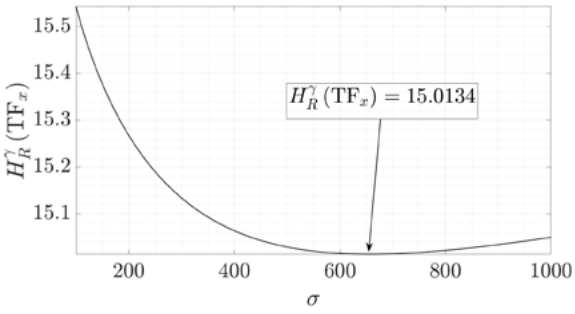
płaszczyźnie czas-częstotliwość $\mathrm{TF}_x(t,\omega)$ jest wyrażona w bitach i zdefiniowana następująco:

$$H_R^{\gamma}(\mathrm{TF}_x) = \frac{1}{1-\gamma}\log_2\left(\frac{\iint_{\mathbb{R}^2}\left(|\mathrm{TF}_x(t,\omega)|\right)^{\gamma}\,\mathrm{d}t\mathrm{d}\omega}{\iint_{\mathbb{R}^2}\left(|\mathrm{TF}_x(t,\omega)|\right)\,\mathrm{d}t\mathrm{d}\omega}\right). \tag{11}$$

W tej pracy $\mathrm{TF}_x(t,\omega)$ odnosi się do (1) i w eksperymencie przyjęto $\gamma = 3$. Im niższa jest wartość entropii, tym bardziej skoncentrowany jest spektrogram. W celu znalezienia optymalnego okna zapewniającego najlepszą rozdzielczość dla danego sygnału, zmieniano parametr σ i obliczano wartość entropii Rényiego (11). Dla wartości entropii osiągającej minimum określono optymalny współczynnik szerokości okna σ. Wyniki przedstawiono na Rysunku 2 (σ wyrażona w jednostkach unormowanych). Wartość optymalna została wyznaczona dla $\sigma = 656$.

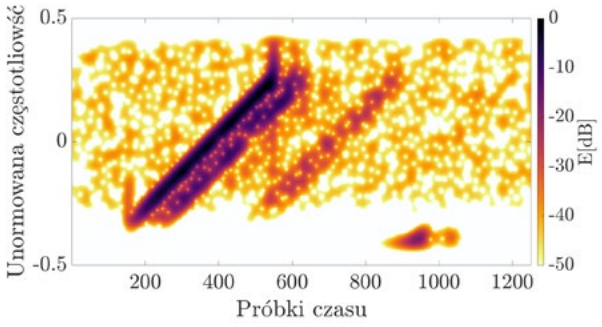


Rysunek 1: Badany sygnał radarowy – przebieg w dziedzinie czasu.

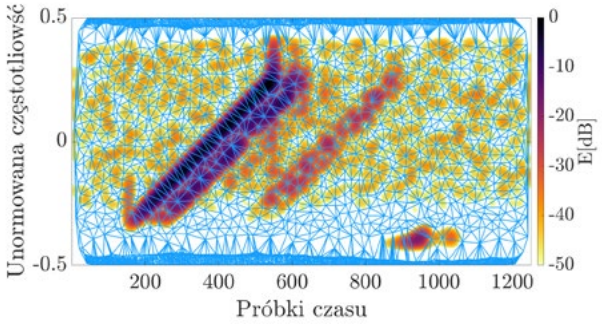


Rysunek 2: Wynik optymalizacji okna analizy – zależność entropii Rényiego od szerokości okna σ.

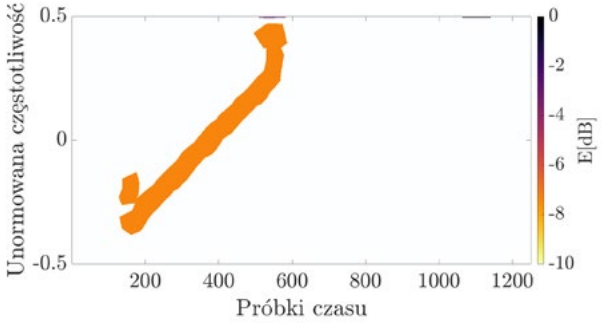
Dla wyznaczonego okna obliczono spektrogram, który pokazano na Rysunku 3. Rozkład sygnału na płaszczyźnie czas-częstotliwość ukazuje dodatkowe składowe, przede wszystkim efekty związane z propagacją wielodrogową, które widoczne są jako opóźnione kopie impulsu. Spektrogram wraz z naniesionym wynikiem triangulacji zer przedstawiono na Rysunku 4. Zgodnie z założeniami algorytmu można zauważyć, że długości boków trójkątów utworzonych przez połączenie zer różnią się w zależności od lokalnego charakteru rozkładu. Trójkąty spełniające warunek długości boku połączono i posortowano względem energii zawartej w każdym wykrytym komponencie. Celem analizy była ekstrakcja najsilniejszego (bezpośredniego) impulsu, zatem zbyt słabe komponenty (np. na skrajach pasma obserwacji) zostały wykluczone z dalszych kroków algorytmu. Na Rysunku 5 pokazano wykryty komponent odzwierciedlający jednocześnie kształt maski binarnej, zgodnie z którą dokonano rekonstrukcji impulsu. Spektrogram po zastosowaniu maski pokazano na Rysunku 6, natomiast porównanie części rzeczywistej i urojonej impulsu przed ($x(t)$) i po filtracji ($\hat{x}(t)$) zaprezentowano odpowiednio na Rysunkach 7 i 8.



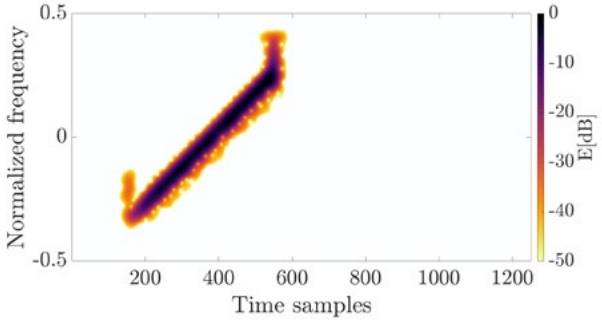
Rysunek 3: Badany sygnał radarowy – spektrogram.



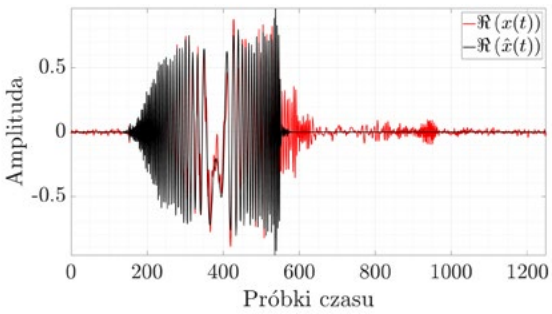
Rysunek 4: Badany sygnał radarowy – spektrogram wraz z wynikiem triangulacji zer.



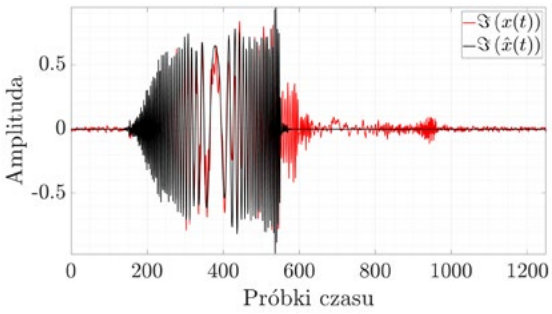
Rysunek 5: Maska dla wykrytego komponentu.



Rysunek 6: Spektrogram po zastosowaniu maski.



Rysunek 7: Porównanie przebiegów czasowych dla sygnału wejściowego ($x(t)$) oraz odfiltrowanego ($\hat{x}(t)$) – część rzeczywista.



Rysunek 8: Porównanie przebiegów czasowych dla sygnału wejściowego ($x(t)$) oraz odfiltrowanego ($\hat{x}(t)$) – część urojona.

Jak można zauważyć, zrekonstruowany impuls został pozbawiony szumu oraz zakłóceń wynikających z propagacji wielodrogowej. Obliczenie odwrotnej transformacji STFT pozwoliło na doknanie rekonstrukcji impulsu z zachowaniem zależności amplitudowych i fazowych i uzyskanie czystej sygnatury radardowej, pomimo braku znajomości parametrów radaru. Potwierdza to skuteczność tej metody w analizie sygnałów niekooperujących nadajników radarowych.

4. PODSUMOWANIE

W niniejszej pracy przedstawiono wyniki dekompozycji sygnału radiolokacyjnego za pomocą metody opartej o triangulację zer spektrogramu. Zaprezentowano podstawy teoretyczne opisujące wykorzystany algorytm, a następnie dokonano filtracji impulsu radarowego pochodzącego z niekooperującego nadajnika radiolokacyjnego. Niniejszy algorytm może posłużyć w aplikacjach związanych z walką radioelektroniczną, a także w radarach pasywnych wykorzystujących pozahoryzontalne nadajniki radarowe do detekcji obiektów latających w systemach wczesnego ostrzegania.

LITERATURA

[1] D. Fourer, F. Auger, K. Czarnecki, S. Meignen, and P. Flandrin. Chirp rate and instantaneous frequency estimation: Application to recursive vertical synchrosqueezing. *IEEE Signal Processing Letters*, 24(11):1724–1728, 2017.

[2] D. Pham and S. Meignen. High-order synchrosqueezing transform for multicomponent signals analysis—with an application to gravitational-wave signal. *IEEE Transactions on Signal Processing*, 65(12):3168–3178, 2017.

[3] P. Flandrin. *Explorations in Time-Frequency Analysis*. Cambridge University Press, 2018.

[4] P. Flandrin. Time–frequency filtering based on spectrogram zeros. *IEEE Signal Processing Letters*, 22(11):2137–2141, 2015.

[5] K. Abratkiewicz and J. Gambrych. Real-Time Variants of Vertical Synchrosqueezing: Application to Radar Remote Sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:1760–1774, 2022.

[6] F. Fioranelli, M. Ritchie, and H. Griffiths. Classification of unarmed/armed personnel using the NetRAD multistatic radar for micro-doppler and singular value decomposition features. *IEEE Geoscience and Remote Sensing Letters*, 12(9):1933–1937, 2015.

[7] V. Bargmann. On a Hilbert space of analytic functions and an associated integral transform part I. *Communications on Pure and Applied Mathematics*, 14(3):187–214, 1961.

[8] T.P. Zieliński. Joint time-frequency resolution of signal analysis using Gabor transform. *IEEE Transactions on Instrumentation and Measurement*, 50(5):1436–1444, 2001.

[9] S. Meignen, M. Colominas, and D. Pham. On the Use of Rényi Entropy for Optimal Window Size Computation in the Short-Time Fourier Transform. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5830–5834, 2020.

[10] R.G. Baraniuk, P. Flandrin, A.J.E.M. Janssen, and O.J.J. Michel. Measuring time-frequency information content using the Rényi entropies. *IEEE Transactions on Information Theory*, 47(4):1391–1409, 2001.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

ROZPROSZONE METODY UKRYWANIA INFORMACJI W SIECIACH

DISTRIBUTED METHODS OF INFORMATION HIDING IN NETWORKS

Jędrzej Bieniasz[1]; Krzysztof Szczypiorski[1]

[1]Instytut Telekomunikacji, Politechnika Warszawska, Warszawa; Jedrzej.Bieniasz@pw.edu.pl

DOI: 10.15199/59.2023.4.28

Streszczenie: Niniejszy artykuł przedstawia rezultaty badań teoretycznych i praktycznych dotyczących różnych aspektów rozproszonych metod ukrywania informacji w sieciach. W ramach prac badawczych podjęto zagadnienie rozproszonych metod ukrywania informacji w sieciach w czterech kierunkach: 1) zbadanie stosowania metod rozproszonej steganografii w ofensywnych operacjach w cyberprzestrzeni, 2) rozszerzenie metodyk modelowania cyberzagrożeń uwzględniając technik ukrywania informacji w sieciach, 3) zrealizowanie badań wykrywania metod steganografii rozproszonej, 4) zastosowanie technik ukrywania informacji jako mechanizmu bezpieczeństwa sieci.
Artykuł opracowano na podstawie rozprawy doktorskiej J. Bieniasza.

Abstract: This article presents the results of theoretical and practical research on various aspects of distributed methods of hiding information in networks. As part of the research work, the issue of distributed information hiding methods in networks was addressed in four directions: 1) testing the use of distributed steganography methods in offensive operations in cyberspace, 2) extending cyber threat modeling methodologies, taking into account information hiding techniques in networks, 3) conducting research on detection of distributed steganography methods, 4) application of information hiding techniques as a network security mechanism.
The article was prepared on the basis of J. Bieniasz's doctoral dissertation.

Słowa kluczowe: steganografia sieciowa, ukrywanie informacji w sieciach, bezpieczeństwo sieci, systemy wieloagentowe, wykrywanie włamań do sieci

Keywords: network steganography, information hiding, network security, multiagent systems, network instrusion detection

1. WSTĘP

Niniejszy artykuł opracowano na podstawie rozprawy doktorskiej J.Bieniasza, na którą składał się cykl 7 publikacji:

[**A1**] **Bieniasz, J.**, & Szczypiorski, K. (2017). **SocialStegDisc: Application of steganography in social networks to create a file system**. 2017 3rd International Conference on Frontiers of Signal Processing (ICFSP), 2017, pp. 76-80. DOI: https://doi.org/10.1109/ICFSP.2017.8097145.

[**A2**] **Bieniasz, J.**, & Szczypiorski, K. (2019). **Methods for Information Hiding in Open Social Networks**. Journal of Universal Computer Science, 25(2), 74-97. DOI: https://doi.org/10.3217/jucs-025-02-0074.

[**A3**] **Bieniasz, J.**, & Szczypiorski, K. (2019). **Steganography Techniques for Command and Control (C2) Channels.** In Botnets. Architectures, Countermeasures, and Challenges. (pp. 189-216). CRC Press. DOI: https://doi.org/10.1201/9780429329913-5.

[**A4**] **Bieniasz, J.**, Stępkowska, M., Janicki, A., & Szczypiorski, K. (2019). **Mobile Agents for Detecting Network Attacks Using Timing Covert Channels.** Journal of Universal Computer Science, 25(9), 1109-1130. DOI: https://doi.org/10.3217/jucs-025-09-1109.

[**A5**] **Bieniasz, J.**, & Szczypiorski, K. (2021). **Dataset Generation for Development of Multi-Node Cyber Threat Detection Systems.** Electronics. 2021; 10(21):2711. DOI: https://doi.org/10.3390/electronics10212711.

[**A6**] **Bieniasz, J.**, & Szczypiorski, K. (2018). **Towards Empowering Cyber Attack Resiliency Using Steganography**. 2018 4th International Conference on Frontiers of Signal Processing (ICFSP), 2018, pp. 24-28. DOI: https://doi.org/10.1109/ICFSP.2018.8552068.

[**A7**] **Bieniasz, J.**, Bąk, P., & Szczypiorski, K. (2022). **StegFog: Distributed Steganography Applied To Cyber Resiliency In Multi Node Environments**. IEEE Access, 2022. DOI: https://doi.org/10.1109/ACCESS.2022.3199749.

Steganografia, w tym steganografia sieciowa i rozproszona, klasyfikowana w domenie cyberbezpieczeństwa w obszarze technik *cyber deception* [1] [2], w ciągu ostatnich 20 lat była intensywnie badana pod kątem możliwości stosowania w działaniach ofensywnych w cyberprzestrzeni, co podsumowuje m.in. publikacja [3]. Rosnącą skalę jej wykorzystania przez atakujących potwierdzają analizy raportów z incydentów bezpieczeństwa komputerowego, które podsumowuje artykuł [**A3**]. Ponadto w artykule tym przedstawiono sposób zastosowania metodyk modelowania cyberzagrożeń do metod steganografii. Samą naturę rozproszonych metod steganografii badano w ramach prac, których wyniki przedstawiono w artykułach [**A1**] i [**A2**]. W artykułach [**A4**] i [**A5**] opisano realizację *end-to-end* procesu badań danych na potrzeby detekcji

metod steganografii rozproszonej – od stworzenia symulacji metod steganografii, przez implementację prototypu systemu detekcji, po referencyjne zastosowanie metod *data science*. W artykule [**A6**] zwrócono uwagę na drugą naturę steganografii, która może stanowić z powodzeniem mechanizm bezpieczeństwa systemów i użytkowników, np. w zakresie zapewniania prywatności czy bezpieczeństwa danych. Artykuł [**A6**] formułuje teoretyczne założenia nowego protokołu dla rozproszonego systemu steganograficznego poprzez zastosowanie koncepcji zaprezentowanych w artykułach [**A1**] i [**A2**]. Artykuł [**A7**] rozwija koncepcję z [**A6**] poprzez implementację prototypu nowego systemu steganograficznego StegFog, przeprowadzenie badań w zakresie parametrów operacyjnych tego systemu oraz dokonanie oceny jego poziomu bezpieczeństwa.

Struktura niniejszego artykułu jest następująca:

- W rozdziale 2 podsumowano badania w zakresie opracowania, implementacji i testów metody steganograficznej, wykorzystującej jako nośnik ukrytych informacji otwarte sieci społecznościowe (OSN). (► zakres artykułów [**A1**] i [**A2**])
- Rodział 3.1 prezentuje badania teoretyczne w zakresie wykorzystania modelowania cyberzagrożeń w procesie projektowania rozwiązań ich wykrywania, ze szczególną koncentracją na metodach ukrywania informacji w sieciach. (► zakres artykułu [**A3**])
- Rozdział 3.2 przedstawia przekrojowe badania nad wykrywaniem rozproszonych metod ukrywania informacji w sieciach z wykorzystaniem podejścia wieloagentowego i teorii obserwacji zmiany. Ważnym aspektem jest też pozyskiwanie źródeł danych do tego typu badań. (► zakres artykułów [**A4**] i [**A5**])
- W rozdziale 4 omówiono madania nad zastosowaniem metod ukrywania informacji w sieciach jako mechanizmu bezpieczeństwa sieci, m.in. proponując nową metodę w tym obszarze – *StegFog*. (► zakres artykułów [**A6**] i [**A7**])
- Rozdział 5 podsumowuje artykuł.

2. REALIZACJA ROZPROSZONYCH METOD UKRYWANIA INFORMACJI W SIECIACH

Niniejszy rozdział podsumowuje badania w zakresie opracowania, implementacji i testów metody steganograficznej, wykorzystującej jako nośnik ukrytych informacji otwarte sieci społecznościowe (OSN). Oryginalna idea została przedstawiona przez K. Szczypiorskiego w [4], a następnie była rozwijana w ramach wspólnych badań.

Jako rozwinięcie metody StegHash autor rozprawy zaproponowano metodę SocialStegDisc w artykule [**A1**]. Pełne rozwinięcie koncepcji, badania i ogólna synteza rozważań nad rozproszonym ukrywaniem informacji w otwartych sieciach społecznościowych została zaprezentowana w artykule [**A2**].

Główne osiągnięcia badawcze zaprezentowane w artykułach [**A1**] i [**A2**] obejmują:

- Opracowanie koncepcji steganograficznego systemu plików o nazwie SocialStegDisc, w którym mechanizm indeksacji danych oparty na StegHash[4] został powiązany z klasyczną konstrukcją systemu plików - listą połączeniową i wskaźnikami na kolejne bloki pamięci. Koncepcja obejmuje dwie możliwości ich utworzenia.
- Zdefiniowanie steganograficznych realizacji typowych operacji klasycznego systemu plików - tworzenie / odczyt / usuwanie.
- Implementacja *proof-of-concept* systemu SocialStegDisc.
- Przeprowadzenie analizy i testów systemu SocialStegDisc w zakresie możliwości skalowania rozmiarów dysku oraz plików, sposobu tworzenia sekretów, wydajności działania w realnym środowisku otwartej sieci społecznościowej.
- Dokonanie ewaluacji aspektów niewykrywalności i niezawodności systemu.
- Wskazanie na problemy szczególne, np. nierównomierność przestrzeni adresowej.
- Przeprowadzenie oceny wpływu wykorzystania takiej koncepcji w cyberprzestrzeni, w szczególności pod kątem jej wykrycia czy też przeanalizowania w ramach procesów kryminalistyki śledczej.
- Powiązanie koncepcji rozproszonych systemów steganograficznych z możliwością ich wykorzystania jako mechanizmu cyberodporności systemów – w odniesieniu do koncepcji zaproponowanej przez Kotta w [5].

3. WYKRYWANIE ROZPROSZONYCH METOD UKRYWANIA INFORMACJI W SIECIACH

3.1 Modelowanie metod steganografii sieciowej

Ten rozdział prezentuje badania teoretyczne w zakresie wykorzystania modelowania cyberzagrożeń w procesie projektowania rozwiązań ich wykrywania. W tym kontekście dokonano przeglądu stanu wiedzy dotyczącej zastosowania steganografii do realizacji komunikacji Command & Control (C2). Ukryta komunikacja jest coraz częściej wykorzystywana przez złośliwe oprogramowanie i botnety oraz w ramach celowanych ataków typu APT (ang. *Advanced Persistent Threats*).

Rezultaty badań teoretycznych autora ujęto w artykule [**A3**]. Na podstawie przeglądu raportów dotyczących rzeczywistych cyberataków i wykrytych botnetów zaproponowano taksonomię podstawowych modeli komunikacji steganograficznych kanałów C2. Przegląd ten objął zarówno analizę złośliwego oprogramowania w tradycyjnych środowiskach komputerowych, jak i platformach mobilnych, dla których istotność zagadnień cyberbezpieczeństwa w ostatnich latach gwałtownie rośnie. Zgodnie z proponowaną strategią cyberobrony opartej na proaktywnym modelowaniu cyberzagrożeń rezultaty artykułu [**A3**] mogą służyć jako *know-how* niezbędne do analizy, wykrywania i przełamywania kanałów C2, gdy są one zabezpieczone steganografią.

Przegląd przedstawiony w artykule [**A3**] obejmował przykłady cyberataków wykorzystujących steganografię do zabezpieczenia kanałów C2 z lat 2010–2018. Po 2018 roku pojawiły się kolejne ataki, które dodatkowo podkreślają istotność poruszanego zagadnienia wykorzystania różnych metod steganografii rozproszonej do zabezpieczenia kanałów C2, np. [6], [7] czy [8]. W momencie powstawania artykułu [**A3**] matryce MITRE ATT&CK [9] nie zawierały technik steganografii. Dlatego w artykule [**A3**]

autor dokonał powiązania analizowanych przykładów kanałów C2 wykorzystujących metody steganografii do technik zawartych w matrycach MITRE ATT&CK. W 2020 roku dokonano oficjalnej aktualizacji matryc, dodając steganografię jako podtechnikę T1001.002 w ramach techniki Obfuskacja danych [T1001] [10]. Pokrywa ona jedynie metody steganografii obrazkowej, zatem mapowanie metod steganografii sieciowej i hybrydowej na matryce MITRE ATT&CK dokonane w artykule [**A3**] stanowi nadal istotny wkład w modelowanie cyberzagrożeń, wykorzystujących te rodzaje metod ukrywania informacji.

3.2 Monitorowanie i wykrywanie kanałów steganograficznych

W tym rozdziale przedstawiono badania w zakresie metod monitorowania i detekcji kanałów steganograficznych w rozproszonych środowiskach teleinformatycznych z wykorzystaniem systemów wieloagentowych. Rezultaty tych badań ujęto w artykułach [**A4**] oraz [**A5**].

Artykuł [**A4**] definiuje koncepcję teorii obserwacji zmiany. Na jej podstawie proponuje algorytm wykrywania anomalii jako niezbędnego kroku w procesie śledzenia ścieżki realizacji ataku, prowadząc w kierunku węzłów źródłowych. Reprezentacją ataków poddawanych monitorowaniu i wykrywaniu są metody steganograficzne oparte na kodowaniu informacji w zależnościach czasowych. W ramach stosowania metody zaprezentowanej w artykule można np. wykryć i śledzić kradzież danych w ramach komunikacji C2. Wykrycie węzłów źródłowych takiej operacji umożliwia efektywne reagowanie na nie, np. przez blokadę ruchu sieciowego.

Realizacją operacyjną tej koncepcji było zastosowanie modelu systemów wieloagentowych (ang. *multi-agent systems*). Systemy cyberobrony oparte na systemach wieloagentowych są analizowane od dawna. Jednym z kamieni milowych rozważań nad tą koncepcją jest metoda przedstawiona w [11]. Kolejne podsumowania prac badawczych w tym obszarze przedstawiono w publikacjach [12], [13] oraz [14]. Ze względu na rozwój środowisk teleinformatycznych i mocy obliczeniowych w ostatniej dekadzie widoczne jest coraz szybsze przyjmowanie się tych koncepcji w rozwiązaniach praktycznych, a zainteresowanie tymi rozwiązaniami w środowiskach akademickich, przemysłu, organów ścigania, a nawet wojska szybko rośnie. Systemy wykrywania cyberzagrożeń oparte na podejściu wieloagentowym są uznawane za skuteczne, a przede wszystkim jako takie, które mogą być jedynymi adekwatnymi i skutecznymi rozwiązaniami w obliczu ewolucji cyberataków, złośliwego oprogramowania i ofensywnych operacji w cyberprzestrzeni.

W artykule [**A4**] opracowano koncepcję systemu wykorzystującego agenty monitorujące w węzłach sieciowych (*routerach programowalnych*), które co określony czas przesyłają stan obserwacji do centralnej jednostki obliczeniowej. Modelem reprezentacji danych przetwarzanych przez agenty i jednostkę centralną są histogramy. W metodzie opisanej w artykule [**A4**] do wykrywania anomalii wykorzystywana jest reprezentacja histogramowa czasu nadejścia pakietów (ang. *Inter Arrival Time* – IAT) [15]. To samo podejście może być z powodzeniem stosowane do obserwacji różnych parametrów, dla których – poza zbieraniem jej atomowej wartości – konieczne jest przechowywanie informacji o stanie histogramów do detekcji cyberzagrożeń, co pokazano m.in. w [16].

Proponowaną w artykule [**A4**] architekturę systemu rozszerza artykuł [**A5**], gdzie zaprezentowano referencyjną architekturę fizyczną i logiczną, obejmującą jako fundamentalne elementy – jednostkę centralną, agenta, router programowalny oraz platformę wieloagentową. Zaimplementowano prototyp platformy wieloagentowej do zarządzania agentami programowymi instalowanymi na routerach z otwartym systemem operacyjnym. Celem działania wskazanych agentów jest przechwytywanie ruchu sieciowego z interfejsów routera, kolekcjonowanie danych do celów wykrywania anomalii oraz śledzenie źródeł ataków sieciowych. Równolegle do prototypu systemu rozwijano środowisko testowe, które w wersji zaprezentowanej w artykule [**A5**] umożliwia:

- szybkie prototypowanie nowych przypadków użycia,
- automatyczne generowanie nowych topologii sieci, tj. decyzji dotyczących:
  - liczby routerów i hostów,
  - zakresu adresów IP i routing,
  - zapewnienia dostępu do Internetu,
  - dołączenia firewalla,
  - umieszczenia jednostki centralnej całego systemu.

Na podstawie tych parametrów opracowane narzędzie generuje losową topologię sieci teleinformatycznej i następnie przesyła ją do silnika emulacji sieci za pośrednictwem interfejsu API lub pliku konfiguracyjnego. Przygotowanie takiej automatyzacji ma na celu zminimalizowanie wpływu topologii sieci na skuteczność nowo opracowanych algorytmów wykrywania cyberzagrożeń poprzez ich trenowanie na danych pozyskanych z różnych topologii. Należy zauważyć, że kluczowym aspektem przygotowanego rozwiązania było zautomatyzowanie losowości mechanizmów:

- generowania łączy pomiędzy zestawem węzłów,
- wyboru nadawców i odbiorców dla każdego profilu strumienia danych sieciowych (normalny lub złośliwy).

Takie podejście pozwala na generowanie zestawów danych z wielu scenariuszy sieciowych i konfiguracji. Może również zapewnić możliwość zminimalizowania wpływu czynników związanych z konfiguracjami w szerokim spektrum badań dotyczących wykrywania cyberzagrożeń na podstawie danych. Zbiory danych zostały zebrane w określonych scenariuszach sieciowych z niewielkim stopniem zmienności w konfiguracji danych sieci nadawca-odbiorca (normalnej lub złośliwej). Potrzeba prac nad takim rozwiązaniem wynika z założenia, że do stosowania badań opartych na metodach analizy danych (ang. *data science*) niezbędne jest posiadanie dobrze przygotowanych zbiorów danych. Metodologia badań zaprezentowana w artykułach [**A4**] i [**A5**] opiera się na samodzielnym pozyskiwania danych na potrzeby badań metod wykrywania, ponieważ wśród przedstawionych publicznych zestawów danych żaden nie oferuje przykładów metod staganografii sieciowej.

W artykule [**A4**] przedstawiono wyniki symulacji metody steganografii kodującej informacje w zależnościach między czasami nadejścia pakietów. Artykuł [**A5**] rozszerzył bazę symulacji o metodę opartą na celowym opóźnianiu i traceniu pakietów. Ponadto obok symulacji metod steganografii możliwe było uruchamianie scenariuszy ruchu sieciowego typowego dla sieci LAN generowanego przy przeglądaniu stron WWW (protokoły

HTTP/HTTPS), komunikacji VoIP (SIP), oglądaniu materiałów wideo (RTP), przesyłaniu plików (FTP/SFTP) czy zdalnym dostępie do serwerów (SSH). Dzięki pozyskaniu właściwych danych możliwe było przeprowadzenie badań nad algorytmami detekcji cyberzagrożeń z wykorzystaniem metod uczenia maszynowego do wykrywania cyberzagrożeń [17].

Typowy proces analizy polega na uczeniu klasyfikatora opartego na uczeniu maszynowym przy użyciu danych, zwykle oznaczonych, uzyskanych ze środowisk produkcyjnych lub emulowanych. Następnie klasyfikator jest testowany przy użyciu danych ewaluacyjnych. Naukowcy zaproponowali wiele różnych algorytmów uczenia maszynowego, począwszy od podstawowych, takich jak k-Nearest Neighbors (k-NN), a skończywszy na algorytmach genetycznych lub różnych typach sieci neuronowych, takich jak perceptrony wielowarstwowe (MLP), mapy Kohonena [18] lub techniki głębokiego uczenia [19].

Badania zaprezentowane w artykule **[A4]** skupiają się na wykorzystaniu metod uczenia maszynowego i mierzonych poziomów klasyfikacji ruchu sieciowego jako złośliwy (klasyfikacje prawdziwie dodatnie (ang. *true positive* – TP) i fałszywie dodatnie (ang. *false positive* – FP) oraz do obliczania metryki anomalii d w zadanym oknie czasowym. Dane pozyskiwane z rozproszonych węzłów są na bieżąco agregowane w rozpatrywanym oknie czasowym i dla każdej z porcji danych ustalana jest wartość metryki d. Przy przekroczeniu określonych progów wartości metryki d oraz w odniesieniu do wartości tej metryki z innych węzłów, algorytm decyzyjny wirtualnie podąża od węzła do węzła w kierunku rosnącej wartości d, osiągając w ten sposób potencjalne źródło ataku.

Badania w artykule **[A4]** w kilku eksperymentach potwierdziły skuteczność proponowanej metody, oferując efektywny sposób na odnajdywanie źródeł cyberataków wykorzystujących steganografię w danej sieci. Najlepszymi algorytmami uczenia w zastosowaniu do wyznaczania metryki d okazały się algorytm lasu losowego (ang. *Random Forest*) oraz klasyczna sztuczna sieć neuronowa (ang. *Multilayer Perceptron*).

Artykuł **[A5]** uzupełnia profil badań z artykułu **[A4]** poprzez:

- przeprowadzenie badań wykrywania cyberzagrożeń na podstawie popularnej reprezentacji danych - modelu strumieniowym, który zawierał 79 różnych metryk statystycznych generowanych przez narzędzie CICFlowMeter [20];
- rozbudowanie procesu analizy danych o:
  - wyrównanie reprezentacji danych próbek złośliwych i normalnych;
  - selekcję istotnych parametrów z wykorzystaniem korelacji;
- przeprowadzenie badań nad metodami uczenia maszynowego w zakresie poprawy jakości klasyfikacji wynikającej z dodania wyżej wymienionych etapów;
- przeprowadzenie badań skuteczności wykorzystania metod uczenia maszynowego w problemie klasyfikacji wieloklasowej, tj. dyskryminacja ruchu normalnego lub jednej z dwóch metod steganografii;
- zbadanie efektywności klasyfikacji cyberataków z wykorzystaniem złożonej architektury sieci neuronowej MLP.

Oryginalnym aspektem badań zaprezentowanych w artykułach **[A4]** i **[A5]** było wykazanie możliwości stosowania technik ukrywania informacji należących do szerokiej gałęzi cyberbezpieczeństwa, określanej jako *cyber deception*. Najważniejszym warunkiem do prowadzenia badań nad wykrywaniem cyberzagrożeń jest dostępność odpowiednich danych. Wszystkie znane zestawy danych, które są dostępne, skupiają się na powszechnie znanych typach cyberataków i nie ma wśród nich przykładów metod steganograficznych. Istotnym krokiem do osiągnięcia rezultatów badawczych było opracowanie środowiska sieciowego zintegrowanego z narzędziami do generowania scenariuszy ataków i następnie zbierania zestawów danych. Osiągnięto to dzięki zaproponowaniu i zaimplementowaniu prototypu systemu wieloagentowego jako wielowęzłowej platformy detekcji anomalii sieciowych. Na podstawie zebranych danych możliwe było przeprowadzenie badań nad algorytmami detekcji wykorzystujących metody reprezentacji danych i klasyfikacji szkodliwych przepływów sieciowych.

Ostateczne wyniki obu artykułów potwierdzają przydatność zbadanych metod detekcji technik ukrywania informacji. Cyberatakujący coraz częściej wykorzystują różne metody z obszaru *cyber deception* w swoich operacjach i w budowaniu wykorzystywanych narzędzi, czyli złośliwego oprogramowania. Wspomniane wcześniej zaawansowane cyberataki APT mogą łączyć wiele technik z różnych warstw systemów teleinformatycznych w każdym z kroków cyberataku modelowanego jako Cyber Kill Chain [21]. W szczególności zastosowanie metod w warstwie aplikacji jest uważane za istotne zagrożenie, co badano bezpośrednio w ramach artykułów **[A1]** i **[A2]**) oraz rozważano teoretycznie w artykule **[A3]**.

4. DEFENSYWNE ZASTOSOWANIA ROZPROSZONYCH METOD UKRYWANIA INFORMACJI W SIECIACH

Niniejszy rozdział omawia badania w zakresie koncepcji rozproszonego ukrywania informacji w sieciach z punktu widzenia zapewniania bezpieczeństwa sieci, cyberodporności i aspektów prywatności. Wpisuje się to w ogólny trend wykorzystania technik z obszaru *cyber deception* w szerszym zakresie rozwiązań w cyberbezpieczeństwie. Techniki *cyber deception* rozważane są od wielu lat także jako rozwiązania defensywne dla cyberobrony, jako np. *honeypoty* [22]. Taksonomię technik *cyber deception* stosowanych w cyberobronie przedstawiają m.in. publikacje [1] oraz [2]. Rezultaty badań autora w zakresie defensywnych zastosowań rozproszonych metod ukrywania zaprezentowano w artykułach **[A6]** oraz **[A7]**.

W ostatnich latach rozwija się paradygmat przetwarzania danych w podejściu rozproszonym, określany jako mgła obliczeniowa (ang. *fog computing*) [23]. W [5] autorzy traktują koncepcję mgły obliczeniowej jako mechanizm bezpieczeństwa sieci i danych, określając ją jako *cyberfog*. Podejście to zakłada, że system cechujący się większą niepewnością z punktu widzenia cyberatakujących zapewnia większą odporność na ataki. Jego bezpośrednią realizacją jest dzielenie danych na wiele fragmentów i rozproszenie ich na wielu urządzeniach użytkowników końcowych. Nawet jeśli system mógłby zostać częściowo skompromitowany, przechwycone informacje byłyby bezużyteczne dla atakującego, a jednocześnie byłyby

nadal przydatne dla użytkowników systemu. Autorzy dostrzegli wiele zalet tego podejścia, wskazując przy tym na wyzwania w zakresie złożoności zarządzania danymi i siecią, przepustowości, pamięci masowej, zapotrzebowania na energię baterii, opóźnień w ponownym odtwarzaniu danych i przerywanej łączności. Ostateczne rozwiązania wymagają właściwego kompromisu między dostępnością a poufnością w zależności od zadań i okoliczności operacyjnych.

Opisana koncepcja cybermgły została podjęta w badaniach zaprezentowanych w artykułach [**A6**] i [**A7**] jako rozwinięcie zastosowań metod opisanych w artykułach [**A1**] i [**A2**]. Efektem tych prac jest koncepcja rozproszonego systemu steganograficznego dla urządzeń IoT, która stosuje założenia mechanizmów StegHash i SocialStegDisc. Koncepcja ta obejmuje architekturę warstwową systemu w zakresie: warstwy aplikacyjnej, systemu plików, sektorów pamięci urządzenia oraz zarządzania pamięcią fizyczną. Każda z tych warstw zawiera, obok komponentu realizacji zadań podstawowych, komplementarny komponent steganograficzny właściwy dla danej warstwy. Całościowo system ma realizować protokół rozproszonej komunikacji kanałami steganograficznymi o charakterze hybrydowym, tj. łączącym steganografię komunikacyjną i steganografię przechowywania danych.

Motywacją badań przedstawionych w artykule [**A7**] jako kontynuacja pomysłu z artykułu [**A6**] było zaproponowanie nowego rozproszonego systemu steganograficznego do ochrony sieci zgodnych z podejściem *cyberfog*, jednocześnie rozwiązującego niektóre ze zdefiniowanych wyzwań dla takiego systemu, co zostało opisane w [5], a także w artykule [**A6**]. Docelowy system musi być możliwy do wdrożenia na urządzeniach końcowych użytkowników i oferować akceptowalny kompromis między dostępnością a poufnością danych. Artykuł [**A7**] formułuje nowy system steganografii rozproszonej, określony jako StegFog. StegFog wykorzystuje schemat indeksowania wprowadzony przez StegHash i zapewnia podstawowe operacje, takie jak w SocialStegDisc (obie metody steganograficzne podsumowują artykuły [**A1**] i [**A2**]). Konstrukcja systemu wykorzystuje pamięć urządzeń końcowych użytkowników jako nośnik dla obiektów z ukrytymi danymi. Proponowane są również metody routingu fragmentów danych i znajdowania kolejnych fragmentów w sieci P2P (ang. *Peer-to-Peer*). Analiza stanu wiedzy przeprowadzona w artykule [**A7**] wskazuje na rosnące potrzeby projektowania, wdrażania i walidacji nowych mechanizmów bezpieczeństwa w podejściu *fog computing*.

Artykuł [**A7**] uzupełnia stan wiedzy w zakresie badań nad rozproszoną steganografią o kompleksowe opracowanie nowego systemu steganograficznego StegFog o charakterze hybrydowym, zastosowanego w połączeniu z podejściem *fog computing*. StegFog łączy trzy typy klasycznych podejść do steganografii w jeden system rozproszonej steganografii:

- steganografię opartą na plikach, np. obrazach (steganografia obrazkowa), do przechowywania danych,
- steganografę tekstową do zarządzania wskaźnikami do ukrytych fragmentów danych,
- steganografię sieciową na potrzeby ukrytej komunikacji w systemie.

W ramach artykułu [**A7**] zaimplementowano działający prototyp systemu StegFog, aby następnie poddać go badaniom w zakresie bezpieczeństwa i wydajności. Ważnym aspektem badania bezpieczeństwa była analiza mechanizmu adresacji StegFog, który opiera się na generowaniu i kodowaniu ukrytych informacji przy pomocy łańcuchów znaków alfanumerycznych. Do ewaluacji wykorzystano entropię względną jako popularną metrykę detekcji dla takich metod w powiązaniu z algorytmicznym przetwarzaniem łańcuchów tekstowych oraz języków. W przypadku Stegfog wykazano wysoką odporność tego mechanizmu. Następnie dokonano oceny parametrów operacyjnych zgodnie z wyzwaniami zdefiniowanymi przez [5], w szczególności pod względem dystrybucji fragmentów danych, złożoności zarządzania siecią, szybkością przesyłania danych, zajętości dysku urządzenia końcowego i wykorzystania jego mocy obliczeniowej. Bezpieczeństwo StegFog zapewnia przede wszystkim charakter logicznego połączenia pomiędzy rozproszonymi częściami łańcucha danych. System mógłby zostać właściwie skompromitowany tylko wtedy, gdyby atakujący przejął moduł funkcji generowania sekretów. Bez niego przechwycone części, np. poprzez skompromitowanie jednego lub kilku urządzeń, byłyby bezużyteczne. Zaproponowane rozwiązanie spełnia tym samym podstawowe wymaganie dotyczące podejścia *cyberfog*. Implementację prototypu StegFog można uznać za jeden z pierwszych działających systemów zgodnych z podejściem *cyberfog*, zdefiniowanym przez [5].

5. PODSUMOWANIE

Główne osiągnięcia badawcze zaprezentowane w niniejszym artykule na podstawie cyklu publikacji [**A1**]–[**A7**] są następujące:

[**O1**] Opracowanie koncepcji nowej metody steganograficznej SocialStegDisc i zaimplementowanie jej prototypu na potrzeby badania jej stosowalności, efektywności i cech charakterystycznych, w szczególności w odniesieniu do możliwości jej wykorzystania w cyberatakach.

[**O2**] Przeprowadzenie analizy teoretycznej stanu wiedzy o zastosowaniu steganografii w kanałach Command & Control (C2), opracowanie autorskiej taksonomii modeli operacyjnych takich kanałów oraz rozszerzenie metodyk modelowania cyberzagrożeń w zakresie stosowania w nich technik steganografii.

[**O3**] Zaprojektowanie i zaimplementowanie autorskiego środowiska do symulacji rozproszonych metod steganografii na potrzeby zbierania zestawów danych do ich analizy.

[**O4**] Opracowanie architektury oraz implementacja prototypu systemu wieloagentowego do monitorowania i wykrywania cyberzagrożeń wykorzystujących metody ukrywania informacji.

[**O5**] Zaproponowanie, implementacja i przeprowadzenie badań weryfikujących słuszność koncepcji nowego algorytmu do wykrywania steganografii w podejściu rozproszonym za pomocą systemu wieloagentowego.

[**O6**] Implementacja i przeprowadzenie badań weryfikujących zastosowanie znanych algorytmów uczenia maszynowego (*Random Forest*, *Multilayer Perceptron*) do wykrywania steganografii w podejściu rozproszonym za pomocą systemu wieloagentowego.

[**O7**] Opracowanie sposobu zastosowania metod StegHash i SocialStegDisc jako mechanizmów architektury bezpieczeństwa danych w koncepcji *fog computing* poprzez zaprojektowanie stosu protokołów do realizacji operacyjnej koncepcji oraz wykonanie analizy aspektów jej bezpieczeństwa.

[**O8**] Implementacja, badania empiryczne oraz ocena bezpieczeństwa prototypu systemu StegFog, który powstał poprzez rozwinięcie założeń przedstawionych w artykule [**A6**].

6. INNE UWAGI

Większość prac badawczych przedstawiona w cyklu publikacji [**A1**]–[**A7**] były realizowana w ramach projektów badawczo-rozwojowych dofinansowanych w programie CyberSecIdent Narodowego Centrum Badań i Rozwoju (NCBiR):

- Artykuły [**A2**] oraz [**A7**] - projekt pt. *Zaawansowane Laboratorium Kryminalistyki Śledczej*, 2017–2019. Numer projektu CYBERSECIDENT/369234/I/NCBR/2017.
- Artykuły [**A3**], [**A4**] oraz [**A5**] - projekt pt. *Platforma detekcji anomalii sieciowych (PDAS)*, 2017–2019. Numer projektu CYBERSECIDENT/369532/I/NCBR/2017.

LITERATURA

[1] N. C. Rowe, "A Taxonomy of Deception in Cyberspace," 2006-03. http://hdl.handle.net/10945/35976.

[2] M. H. Almeshekah and E. H. Spafford, *Cyber Security Deception*, pp. 23–50. Cham: Springer International Publishing, 2016.

[3] W. Mazurczyk, S. Wendzel, S. Zander, A. Houmansadr, and K. Szczypiorski, *Information Hiding in Communication Networks: Fundamentals, Mechanisms, Applications, and Countermeasures*. Wiley-IEEE Press, 1st ed., 2016.

[4] K. Szczypiorski, "StegHash: New Method for Information Hiding in Open Social Networks," *IJET International Journal of Electronics and Telecommunications*, vol. 62, no. 4, pp. 347–352, 2016.

[5] A. Kott, A. Swami, and B. West, "The fog of war in cyberspace," tech. rep., US Army Research Laboratory, 2016.

[6] Threatpost, "MoleRats APT Returns with Espionage Play Using Facebook, Dropbox," 2020. https://threatpost.com/molerats-apt-espionage-facebook-dropbox/162162/, [Online; Dostęp: 23.06.2023].

[7] Security Affairs, "njRAT RAT operators leverage Pastebin C2 tunnels to avoid detection," 2020. https://securityaffairs.co/wordpress/112147/cyber-crime/njrat-rat-pastebin-c2.html, [Online; Dostęp: 23.06.2023].

[8] Cisco Talos Intelligence Group, "North Korean attackers use malicious blogs to deliver malware to high-profile South Korean targets," 2021. https://blog.talosintelligence.com/2021/11/kimsuky-abuses-blogs-delivers-malware.html, [Online; Dostęp: 23.06.2023].

[9] "MITRE ATT&CK: globally-accessible knowledge base of adversary tactics and techniques based on real-world observations.." https://attack.mitre.org. [Online; Dostęp: 23.06.2023].

[10] "MITRE ATT&CK: globally-accessible knowledge base of adversary tactics and techniques based on real-world observations. Data Obfuscation: Steganography (T1001.002).." https://attack.mitre.org/techniques/T1001/002/. [Online; Dostęp: 23.06.2023].

[11] J. Balasubramaniyan, J. Garcia-Fernandez, D. Isacoff, E. Spafford, and D. Zamboni, "An architecture for intrusion detection using autonomous agents," in *Proceedings 14th Annual Computer Security Applications Conference (Cat. No.98EX217)*, pp. 13–24, 1998.

[12] A. Herrero and E. Corchado, "Multiagent Systems for Network Intrusion Detection: A Review," in *Advances in Intelligent and Soft Computing*, vol. 63, pp. 143–154, 01 2009.

[13] M. Docking, A. V. Uzunov, C. Fiddyment, R. Brain, S. Hewett, and L. Blucher, "UNISON: Towards a Middleware Architecture for Autonomous Cyber Defence," in *2015 24th Australasian Software Engineering Conference*, pp. 203–212, 2015.

[14] I. A. Saeed, A. Selamat, M. F. Rohani, O. Krejcar, and J. A. Chaudhry, "A Systematic State-of-the-Art Analysis of Multi-Agent Intrusion Detection," *IEEE Access*, vol. 8, pp. 180184–180209, 2020.

[15] A. Kind, M. P. Stoecklin, and X. Dimitropoulos, "Histogram-based traffic anomaly detection," *IEEE Transactions on Network and Service Management*, vol. 6, no. 2, pp. 110–121, 2009.

[16] Z. Miller, W. Deitrick, and W. Hu, "Anomalous network packet detection using data stream mining," *Journal of Information Security*, vol. 2, no. 04, p. 158, 2011.

[17] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016.

[18] C.-F. Tsai, Y.-F. Hsu, C.-Y. Lin, and W.-Y. Lin, "Intrusion detection by machine learning: A review," *Expert Systems with Applications*, vol. 36, no. 10, pp. 11994–12000, 2009.

[19] D. Kwon, H. Kim, J. Kim, S. C. Suh, I. Kim, and K. J. Kim, "A survey of deep learning-based network anomaly detection," *Cluster Computing*, Sep 2017.

[20] "CICFlowMeter: Ethernet traffic Bi-flow generator and analyzer," 2018. https://https://github.com/ahlashkari/CICFlowMeter, [Online; Dostęp: 05.02.2020 (wersja aplikacji)].

[21] E. M. Hutchins, M. J. Cloppert, and R. M. Amin, "Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains," *Leading Issues in Information Warfare and Security Research*, vol. 1, 01 2011.

[22] L. Spitzner, "The Honeynet Project: trapping the hackers," *IEEE Security & Privacy*, vol. 1, no. 2, pp. 15–23, 2003.

[23] M. Chiang and T. Zhang, "Fog and IoT: An Overview of Research Opportunities," *IEEE Internet of Things Journal*, vol. 3, no. 6, pp. 854–864, 2016.

EFEKTYWNA ENERGETYCZNIE ALOKACJA ZASOBÓW W SYSTEMACH OFDM ZE ŚWIADOMOŚCIĄ OBLICZENIOWĄ

ENERGY-EFFICIENT RESOURCE ALLOCATION
IN OFDM SYSTEMS WITH COMPUTATIONAL AWARENESS

Bartosz Bossy
bartoszbossy@gmail.com

DOI: 10.15199/59.2023.4.29

Streszczenie: W artykule przedstawiono nowe metody optymalizacji dla łącza i sieci bezprzewodowej OFDM/OFDMA (ang. Orthogonal Frequency-Division Multiplexing/Orthogonal Frequency-Division Multiple access). W przeciwieństwie do tradycyjnego podejścia np. maksymalizacji przepływności czy tez minimalizacji mocy transmisji, artykuł koncentruje się na optymalizacji metryki efektywności energetycznej. Ponadto, w badaniach zastosowano zaawansowane modele zużycia energii, uwzględniające moc zużywaną przez przetwarzanie sygnału w paśmie podstawowym. Ten składnik może mieć kluczowe znaczenie w przypadku komunikacji na niewielkie odległości, dla której moc zużywana na przetwarzanie sygnału może dominować nad mocą potrzebną do emisji sygnału. Wyniki badań pokazują, że zaproponowane metody zwiększają efektywność energetyczną we wszystkich rozważanych scenariuszach.

Abstract: In this paper, new optimization methods for OFDM/OFDMA link and wireless network are presented. In contrast to the approaches known from the literature, such as maximizing the throughput or minimizing the transmission power, this article focuses on optimizing the energy efficiency metric. Moreover, the advanced power consumption models that consider the power consumed by baseband signal processing are used. This component can be critical in short-distance communications, where the energy used for signal processing may dominate over the power needed to transmit the signal. The research results show that the proposed methods increase energy efficiency in all scenarios.

Słowa kluczowe: alokacja zasobów, efektywność energetyczna, OFDM/OFDMA.

Keywords: energy efficiency, OFDM/OFDMA, resource allocation

1. WSTĘP

W dobie wszechobecnego dostępu do Internetu z roku na rok można zaobserwować wykładniczy wzrost ruchu telekomunikacyjnego. Zgodnie z prognozami Cisco w 2022 r. 4,8 miliarda użytkowników oraz 28,5 miliarda urządzeń jest podłączonych do sieci Internet [1]. Ponadto ruch pochodzący z urządzeń mobilnych wzrósł do 930 eksabajtów w 2022 roku. Jednym z problemów jest wzrost zużycia energii związany ze wzrostem ruchu telekomunikacyjnego na całym świecie. Właśnie dlatego komunikacja bezprzewodowa o zwiększonej efektywności energetycznie (ang.energy efficiency - EE) jest w ostatnich latach w centrum zainteresowania badań i przemysłu, dążąc do osiągnięcia 10-krotnie wyższej EE w systemach radiowych 5. generacji (5G) w porównaniu z systemami 4. generacji (4G). Co więcej, zmniejszenie ogólnego zużycia energii w sieci będzie krytycznym wymogiem dla tych przyszłych systemów bezprzewodowych. Wraz z ukończeniem przez 3GPP wydania 15 i wydania 16, zdefiniowano podstawy systemu 5G. Podobnie jak w przypadku systemu czwartej generacji dla systemów 5G zaproponowano technikę OFDM. Popularność techniki OFDM wynika m.in. z wysokiej wydajności widmowej (ang. spectral efficiency - SE), elastyczności m.in. możliwości dostosowania parametrów do zmieniających się warunków kanału. W przypadku OFDM wielodostęp może być zrealizowany w warstwie MAC z użyciem OFDMA. Motywowany powyższymi aspektami w artykule przedstawiono badania, które doprowadziły do zaproponowania nowych metod optymalizacji dla łącza i sieci bezprzewodowej OFDM/OFDMA. W przeciwieństwie do tradycyjnego podejścia np. maksymalizacji przepływności przy ograniczeniu na zużywaną moc albo minimalizacji mocy transmisji dla założonej docelowej przepływności, artykuł koncentruje się na optymalizacji metryki efektywności energetycznej, definiowanej jako liczba pomyślnie przesłanych i odebranych bitów na dżul. Ponadto, autor stosuje zaawansowane modele zużycia energii, uwzględniające moc zużywaną przez przetwarzanie sygnału w paśmie podstawowym. Ten składnik może mieć kluczowe znaczenie w przypadku komunikacji na niewielkie odległości, dla której moc zużywana na przetwarzanie sygnału może dominować nad mocą potrzebną do emisji sygnału.

Głównym celem artykułu jest przedstawienie zaproponowanych technik maksymalizujących EE, a w szczególności:

- Opracowanie efektywnego, nisko złożonego algorytmu maksymalizacji metryki EE ze świadomością obliczeniową w łączu OFDM z adaptacyjnym doborem mocy transmisji oraz wyborem modulacji i kodowania.
- Opracowanie algorytmu alokacji zasobów maksymalizującego EE w sieci OFDMA z wieloma użytkownikami, z uwzględnieniem ograniczeń praktycznych systemów komunikacji bezprzewodowej.
- Rozwiązanie problemu optymalizacji EE dla sieci przekaźnikowej z wieloma użytkownikami oraz in-

terferencją między nimi z uwzględnieniem mocy zużywanej na przetwarzanie sygnału w paśmie podstawowym.

2. PRZYDZIAŁ ZASOBÓW MAKSYMALIZUJĄCY EE

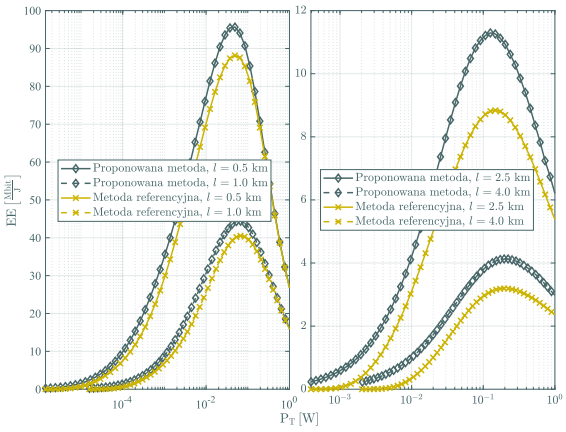
W niniejszym rozdziale przedstawiono główne wyzwania i problemy przydziału zasobów w systemach komunikacji bezprzewodowej maksymalizujących EE. W kontekście systemów bezprzewodowych EE jest definiowana jako stosunek przepustowości osiąganej w systemie do całkowitej mocy zużywanej przez system:

$$\eta \left[\frac{\text{bit}}{\text{dżul}}\right] = \frac{\text{przepływność}\,[\text{bit/s}]}{\text{zużywana moc}\,[\text{W}]}. \tag{1}$$

Oznacza to, że efektywność energetyczna określa liczbę przesłanych bitów przypadających na jednostkę energii. Zatem im wyższa metryka EE tym rozpatrywany system jest bardziej wydajny energetycznie. W artykule zidentyfikowano różne podejścia do maksymalizacji EE oraz dokonano klasyfikacji zadań jakie należy wykonać projektując algorytm alokacji zasobów maksymalizujący EE. Wyodrębniono następujące zadania: (i) estymacja przepustowości, (ii) estymacja całkowitej zużywanej mocy w systemie, (iii) identyfikacja wymagań i ograniczeń systemowych, (iv) optymalizacja metryki EE. Dla każdego z powyższych zadań zidentyfikowano główne podejścia stosowane w literaturze, przedstawiono ich główne wady i zalety oraz zaproponowano kompromis między ich szczegółowością, a złożonością. W kontekście estymacji przepustowości zidentyfikowano następujące podejścia: (i) estymacja z użyciem równania Shannona, (ii) estymacja z użyciem równania Shannona zmodyfikowanym przez współczynniki skalujące, (iii) estymacja z użyciem funkcji błędu i wydajności widmowej dla danego schematu modulacji i kodowania. W przypadku estymacji całkowitej mocy zużywanej w systemie zidentyfikowano następujące podejścia: (i) wysokopoziomowy model zużycia mocy, (ii) estymacja mocy bazująca na pomiarach, (iii) estymacja mocy jako sumy mocy zużywanej przez każdy komponent nadajnik i odbiornika. Należy pamiętać. że dla bardziej szczegółowych modeli (np. z większą liczbą parametrów) problem optymalizacyjny maksymalizujący EE może być trudniejszy do rozwiązania co przekłada się na złożoność obliczeniową algorytmu alokacji zasobów. W kontekście identyfikacji wymagań i ograniczeń systemowych zidentyfikowano najczęściej spotykane wymagania i ograniczenia rozważane w kontekście alokacji zasobów maksymalizujących EE. Ponadto, w dokonano przeglądu oraz porównania metod optymalizacyjnych stosowanych w celu maksymalizacji EE zwracając uwagę na ich wady i zalety.

3. OPTYMALIZACJA EFEKTYWNOŚCI ENERGETYCZNEJ W ŁĄCZU OFDM

W niejszym rozdziale przedstawiono zaproponowany algorytm alokacji zasobów maksymalizujący EE dla łącza OFDM z adaptacyjnym dobrem mocy transmisji oraz schematu modulacji i kodowania [2]. W odróżnieniu od rozwiązań dostępnych w literaturze [3, 4] w rozważanym modelu uwzględniono moc zużywaną na przetwarzanie sygnału w paśmie podstawowym zależne od wybranego schematu modulacji i kodowania, który jest dobierany adaptacyjne w zależności od warunków radiowych oraz wymagań użytkownika np. bitowej stopy błędu. Dla opisanego powyżej modelu systemu zdefiniowano problem optymalizacyjny, który sprowadza się do znalezienia takich wartości mocy transmisji oraz schematów modulacji i kodowania alokowanych na podnośne, które maksymalizują EE przy ograniczeniu na całkowitą maksymalną moc transmisji oraz minimalną osiąganą bitową stopę błędu. Ponieważ metryka EE ma postać ułamkową, dlatego też problem optymalizacyjny jest z natury niewypukły. Dodatkowo, należy pamiętać, że w opisywanym problemie optymalizacyjnym mamy do czynienia z ciągłymi zmiennymi określającymi moc transmitowaną na podnośnych oraz ze zmiennymi dyskretnymi opisującymi parametry stosowanych koderów oraz dekoderów. Tak zdefiniowane problemy optymalizacyjne noszą nazwę ciągłodyskretnych nieliniowych problemów optymalizacyjnych, które są bardzo trudne do rozwiązania w oryginalnej postaci. Dlatego też dokonano przekształcenia oryginalnego problemu optymalizacyjnego, a następnie zaproponowano algorytm alokacji zasobów o niskiej złożoności obliczeniowej maksymalizujący EE. Zaproponowany algorytm alokacji zasobów oparty na algorytmie Dinkelbacha, o niskiej złożoności obliczeniowej, został zweryfikowany przez symulację komputerową. EE w funkcji całkowitej mocy transmisji dla różnych odległości między nadajnikiem, a odbiornikiem l została przedstawiona na rysunku 1. Wyniki uzyskane z wykorzystaniem proponowanego algorytmu zostały porównane z metodą referencyjną, która maksymalizuje przepustowość alokując moc zgodnie z zasadą „wypełniania wodą", a następnie w zależności od wartości stosunku sygnału do szumu wybiera odpowiedni schemat modulacji i kodowania. Wyniki symulacji pokazują, że proponowany algorytm pozwala na zwiększenie EE w porównaniu z metodą referencyjną. Ponadto, można zauważyć, że istnieje optymalny punkt dla mocy nadawania maksymalizujący EE. Zysk w EE proponowanego rozwiązania w porównaniu z metodą referencyjną wynosi od 10% do 15% w zależności od scenariusza, co może skutkować zmniejszeniem zużycia energii w systemach komunikacji bezprzewodowej OFDM.



Rysunek 1: Efektywność energetyczna w funkcji całkowitej mocy transmisji.

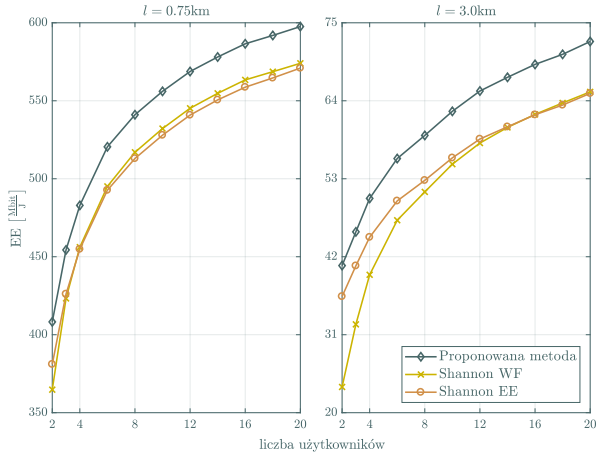
4. EFEKTYWNOŚĆ ENERGETYCZNA DLA SIECI OFDMA

Kolejnym etapem badań była optymalizacja EE dla sieci urządzeń używającej techniki OFDMA. Zaproponowano algorytm przydziału zasobów dla sieci OFDMA z wielo-

ma użytkownikami [5]. W przeciwieństwie do istniejących prac [6, 7] w rozważanym modelu zostały uwzględnione ograniczenia praktycznych systemów bezprzewodowych, w których: (i) podnośne są grupowane w bloki zasobów (ang. resource block - RB), (ii) wszystkie bloki zasobów przypisane do użytkownika są transmitowane z tą samą mocą używając tego samego schematu modulacji i kodowania, (iii) przepustowość wynika z blokowej stopy błędy oraz efektywności widmowej zastosowanego schematu modulacji i kodowania. Ponieważ każdy blok zasobów składa się z pewnej liczby sąsiadujących podnośnych charakteryzujących się różnymi wartościami tłumienia, konieczna jest metoda wyznaczająca efektywną wartość stosunku sygnału do szumu dla całego zbioru podnośnych zaalokowanych dla danego użytkownika. Ponadto, w celu wyznaczenia przepustowości wymagana jest estymacja blokowej stopy błędu dla każdego dostępnego w systemie schematu modulacji i kodowania.

Biorąc pod uwagę cechy opisanego powyżej modelu systemu okazuję się, że zdefiniowany dla niego problem optymalizacyjny maksymalizujący EE nie tylko jest niewypukłym ciągło-dyskretnym problemem optymalizacyjnym, ale również problemem kombinatorycznym. Poza ciągłymi zmiennymi określającymi efektywny SNR mamy do czynienia ze skończonym zbiorem schematów modulacji i kodowania oraz binarnymi zmiennymi określającymi czy blok zasobów jest przypisany do danego użytkownika czy nie. W związku z powyższym oryginalny problem optymalizacyjny został przekształcony poprzez wprowadzenie macierzy decyzyjnej oraz zastosowanie metody Dinkelbacha. Zaproponowano dwuetapowy, iteracyjny algorytm alokacji zasobów, który został zweryfikowany za pomocą symulacji komputerowej i porównany z metodami referencyjnymi.

Analizowany system składał się z 15 schematów MCS zdefiniowanych w specyfikacji systemu LTE (ang. Long-Term Evolution), podobnie jak całkowita liczba bloków zasobów należąca do zbioru $\{6, 15, 25, 50, 75, 100\}$ oraz 12 podnośnych w każdym RB. Przyjęty model kanału to Extended Pedestrian A [**?**]. Pierwszy algorytm referencyjny, nazwany jako *Shannon EE*, z którym porównano opisywaną metodą, bazuje na pracy [**?**] i został zaadaptowany, tak aby liczba stopni swobody była identyczna jak w proponowanej metodzie.



Rysunek 2: Efektywność energetyczna w funkcji liczby liczby użytkowników w sieci dla różnych promieni komórki l.

Drugi algorytm odniesienia, nazwany *Shannon WF*, bazuje na znanej zasadzie "wypełniania wodą". Oba algorytmy referencyjne alokują moc dla poszczególnych użytkowników, w pierwszym przypadku w celu maksymalizacji EE, w drugim w celu maksymalizacji przepływności, lecz nie uwzględniają doboru schematów MCS. Dlatego też schemat MCS jest dobierany na podstawie wartości efektywnego SNR. Przykładowe wyniki pokazano na Rysunku 2 przedstawiającym EE w funkcji liczby użytkowników. Zauważalne jest, że proponowane rozwiązanie osiąga lepsze rezultaty niż metody referencyjne oraz, że EE rośnie wraz z liczbą użytkowników.

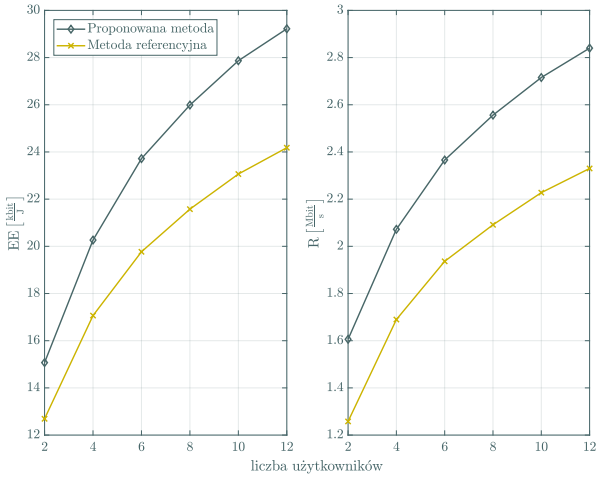
5. OPTYMALIZACJA EE W SIECI OFDMA Z PRZEKAŹNIKAMI

Kolejna część artykułu dotyczy zastosowania węzłów przekaźnikowych w sieci OFDMA w celu zwiększenia EE transmisji. W kontekście EE zastosowanie przekaźników pozwala na zmniejszenie długości łączy, a co za tym idzie zmniejszenie zużywanej mocy dzięki niższej mocy transmisji lub zastosowaniu mniej złożonego obliczeniowo kodowania kanałowego. Z drugiej jednak strony, dodanie przekaźnika zwiększa liczbę urządzeń radiowych, których moc zużywaną trzeba dodać do całkowitego modelu zużycia mocy. Ponadto, transmisja z wykorzystaniem węzła przekaźnikowego wymaga co najmniej dwóch szczelin czasowych.

W artykule przedstawiono zaproponowany algorytm alokacji zasobów dla sieci OFDMA z przekaźnikami działającymi w trybie DF (ang. decode and forward - DF) [9, 10]. W przeciwieństwie do istniejących systemów [11, 12, 13] w proponowanym modelu ta sama podnośna może być wykorzystywana przez stację bazową i przez węzły przekaźnikowe w drugiej szczelinie czasowej. Takie podejście może powodować interferencję u użytkowników końcowych lecz w przypadku gdy tłumienie łącza przez które propaguje się sygnał interferujący względem łącza dla sygnału pożądanego jest duże możliwe jest zwiększenie efektywności widmowej systemu co może skutkować zwiększeniem EE. Ponadto, w rozważanym systemie uwzględniono moc zużywaną na przetwarzanie sygnału w paśmie podstawowym, a tryb transmisji (bezpośredni lub z wykorzystaniem przekaźnika) jest dobierany adaptacyjne w zależności od warunków kanału. Sformułowany dla powyższego systemu problem optymalizacyjny sprowadza się do znalezienia takich wartości mocy transmisji alokowanych na podnośne, par podnośnych alokowanych w pierwszej i drugiej szczelnie czasowej oraz trybów transmisji użytkowników sieci, które maksymalizują EE. Podobnie jak w przypadku łącza OFDM jak i sieci OFDMA zdefiniowany problem optymalizacyjny jest niewypukłym, ciągłodyskretnym problemem optymalizacyjnym. Ponadto, ze względu na niewypukłość licznika metryki EE metody rozwiązania problemów ułamkowych nie mogę zostać użyte. Dlatego też, w celu rozwiązania zdefiniowanego problemu dokonano jego przekształcenia.

Ostatecznie zaproponowano algorytm przydziału zasobów maksymalizujący EE w oparciu o metodę Dinkelbacha, sukcesywną aproksymację funkcją wypukłą oraz algorytm węgierski, który został zweryfikowany poprzez symulację komputerową. Przykładowy wynik symulacji pokazujący efektywność energetyczną oraz przepustowość w funkcji liczby użytkowników w sieci pokazano na rysunku 3. Metoda referencyjna maksymalizuje efektywność energetyczną systemu podobnie jak zaproponowany

algorytm, ale nie może wykorzystywać danej podnośnej równocześnie przez stację bazową i węzeł przekaźnikowy w drugiej szczelnie czasowej. Wyniki symulacji pokazują znaczący zysk zarówno w efektywności energetycznej jak i osiąganej przepływności. Pokazuje to dobitnie zalety optymalizacji sieci z uwzględnieniem interferencji.



Rysunek 3: Efektywność energetyczną oraz przepustowość w funkcji liczby użytkowników w sieci z przekaźnikami.

6. PODSUMOWANIE

Maksymalizacja metryki EE definiowanej nie jest trywialnym problemem. Dzieje się tak, ponieważ zwiększona przepływność w nowoczesnych systemach bezprzewodowych osiągana jest kosztem złożoności obliczeniowej urządzeń radiowych wzrastającej z uwagi na bardziej zaawansowane przetwarzanie sygnałów np. kodowanie, dekodowanie, modulacja. Powoduje to, że energia związana z przetwarzaniem sygnałów, niezależnie od energii transmisji, stanowi znaczną część całkowitego zużycia energii. W związku z tym nowoczesne, energooszczędne systemy i sieci komunikacji radiowej muszą być *świadome obliczeniowo*. Ponadto zróżnicowanie systemów komunikacji bezprzewodowej, różne implementacje sprzętowe stosowanych algorytmów oraz praktyczne ograniczenia tych systemów znacznie komplikują problemy optymalizacyjne. W artykule przedstawiono trzy scenariusze, które zostały zbadane. Autor omawia systemy o narastającym stopniu złożoności, od: (i) pojedynczego łącze OFDM poprzez (ii) sieć OFDMA, a skończywszy na (iii) sieci OFDMA z przekaźnikami. W każdym przypadku należy zmierzyć się z estymację szybkości transmisji, modelowaniem zużycia energii, istniejących ograniczeniach i wymaganiach systemowych. Proponowany jest problem optymalizacyjny oraz jego rozwiązanie. W szczególności w modelowaniu zużycia energii uwzględniona jest moc przetwarzania sygnału (i związaną z tym moc obliczeniową). Ponadto, w każdym z badanych scenariuszy zaproponowano nowatorskie rozwiązania i autorskie algorytmy alokacji zasobów, które pozwalają zwiększyć EE w stosunku do istniejących rozwiązań. Przeprowadzone badania pozwalają stwierdzić, że zaproponowano nowe metody maksymalizacji EE w systemach OFDM/OFDMA ze świadomością obliczeniową.

LITERATURA

[1] Cisco. 2018. „Complete Visual Networking Index (VNI) Complete Forecast Update (2017 – 2022)".

[2] B. Bossy i H. Bogucka. 2016. „Optimization of Energy Efficiency in Computationally-Aware Adaptive OFDM Systems". IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications 2016 (IEEE PIMRC 2016) : 1–6.

[3] G. Miao, N. Himayat i G. Y. Li. 2010. „Energy-Efficient Link Adaptation in Frequency-Selective Channels". IEEE Transac- tions on Communications 58(2): 545–554, lut. 2010.

[4] C. Isheden i G. P. Fettweis. 2010. „Energy-Efficient Multi-Carrier Link Adaptation with Sum Rate- Dependent Circuit Power". IEEE Global Telecommunications Conference 2010 (IEEE GLOBECOM 2010) : 1–6.

[5] B. Bossy, P. Kryszkiewicz i H. Bogucka. 2017. „Optimization of Energy Efficiency in the Downlink LTE Transmission". IEEE International Conference on Communications 2017 (ICC 2017) : 1–6.

[6] A. Akbari, R. Hoshyar i R. Tafazolli, "Energy-efficient resource allocation in wireless OFDMA systems," w PIMRC'11, wrz. 2010, s. 1731–1735. doi: 10.1109/PIMRC.2010.5671913.

[7] C. Xiong, G. Y. Li, S. Zhang, Y. Chen i S. Xu. 2012. „Energy-Efficient Resource Allocation in OFDMA Networks". IEEE Transactions on Communications 60,(12) : 3767–3778.

[8] K. Yang, S. Martin, T. A. Yahiya i J. Wu. 2014. „Energy-Efficient Resource Allocation for Downlink in LTE Heterogeneous Networks". IEEE 80th Vehicular Technology Conference 2014 (VTC2014-Fall) : 1–5.

[9] B. Bossy, P. Kryszkiewicz i H. Bogucka. 2018. „Energy Efficient Resource Allocation in Multiuser DF Relay Interference Networks". IEEE Global Telecommunications Conference Workshops 2018 (IEEE GLOBECOM Wkshps 2018) : 1–6.

[10] B. Bossy, P. Kryszkiewicz i H. Bogucka. 2018. „Energy Efficient Wireless Relay Networks With Computational Awareness". IEEE Transactions on Communications 68(2) : 825–840.

[11] K. T. K. Cheung, S. Yang i L. Hanzo. 2013. „Achieving Maximum Energy-Efficiency in Multi-Relay OFDMA Cellular Networks: A Fractional Programming Approach". IEEE Transactions on Communications 61(7) : 2746–2757.

[12] R. Arab Loodaricheh, S. Mallick i V. K. Bhargava. 2014. „Energy-Efficient Resource Allocation for OFDMA Cellular Networks With User Cooperation and QoS Provisionin". IEEE Transactions on Wireless Communications 13(11) : 6132–6146.

[13] Y. Lu, K. Xiong, Y. Zhang, P. Fan i Z. Zhong. 2016. „Energy-Efficient Resource Allocation in OFDM Relay Networks under Proportional Rate Constraints". IEEE Global Communications Conference 2016 (IEEE GLOBECOM 2016) : 1–6.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

WYKORZYSTANIE INFORMACJI KONTEKSTOWEJ DO POPRAWY EFEKTYWNOŚCI WIDMOWEJ W SYSTEMACH RADIOWYCH

UTILIZATION OF CONTEXT INFORMATION FOR SPECTRAL-EFFICIENCY ENHANCEMENT IN FUTURE RADIO COMMUNICATION SYSTEMS

Łukasz Kułacz[1]

[1] Politechnika Poznańska, Poznań, lukasz.kulacz@put.poznan.pl

DOI: 10.15199/59.2023.4.30

Streszczenie: Wobec wymagań dotyczących zasobów radiowych jednym z rozważanych podejść jest wykorzystanie systemów dynamicznego alokacji zasobów radiowych. Systemy te potrzebują informacji kontekstowej o użytkownikach i samej sieci, aby optymalnie zarządzać widmem. W pracy rozważono wykorzystanie informacji kontekstowej do poprawy efektywności widmowej w sieciach radiowych. Rozwiązania zostały zweryfikowane poprzez symulacje działania sieci, a niektóre zaimplementowano i przetestowano z wykorzystaniem rzeczywistych urządzeń.

Abstract: In view of the radio resource requirements, one of the approaches being considered is the use of dynamic radio resource allocation systems. These systems need contextual information about users and the network itself to optimally manage the spectrum. The paper considers the use of contextual information to improve spectral efficiency in radio networks. The solutions have been verified through simulations of network operation, and some have been implemented and tested using real devices.

Słowa kluczowe: dynamiczny przydział zasobów, efektywność widmowa, informacje kontekstowe

Keywords: context information, dynamic resource allocation, spectral efficiency.

1. WSTĘP

Współcześnie telekomunikacja bezprzewodowa to już nie tylko środek komunikacji między ludźmi, ale bardziej - a może nawet przede wszystkim - wygodny sposób na dostarczanie treści i multimediów niezależnie od tego, gdzie się znajdujemy. Oferowane usługi charakteryzują się różnymi wymaganiami, przeważnie odnoszącymi się do oczekiwanych wartości przepływności. Coraz większa liczba urządzeń końcowych, wyposażonych w wyświetlacze o dużych rozmiarach i o dużej rozdzielczości skutkuje ogromnym wzrostem zapotrzebowania ma ilość przesyłanych danych. Pewnym rozwiązaniem tego problemu jest przydzielanie i wykorzystywanie większej szerokości pasma do transmisji. Jednak statyczne przydzielanie zasobów radiowych różnym usługom powoduje, że swobodne ich wykorzystanie jest niemożliwe, głównie ze względu na umowy, regulacje i ograniczenia licencyjne. Wykorzystanie kolejnych (dotychczas nieużywanych) zakresów częstotliwości z kolei wiąże się z bardziej skomplikowanym środowiskiem propagacyjnym i często z koniecznością projektowania i produkcji nowych urządzeń radiowych. W konsekwencji taki podział zasobów radiowych nie jest efektywny, a w wielu sytuacjach wiąże się z niewystarczającą alokacją zasobów dla jednej usługi i nadmiarem zasobów dla innej. Można jednak przyjrzeć się obecnie zarezerwowanym już fragmentom pasma radiowego i ich wykorzystaniu; w wielu przypadkach zauważyć można obszary geograficzne i określone przedziały czasowe, w których niektóre fragmenty pasma nie są wykorzystywane. W świetle ogromnego zapotrzebowania na widmo radiowe podejście to wymaga przemyślenia. Ponieważ popyt na usługi bezprzewodowe w większości przypadków jest dynamiczny i zależy od różnych czynników (takich jak pora dnia, dzień tygodnia, pora roku, ale także lokalizacja), podział zasobów radiowych powinien być również dynamiczny.

Efektywność widmowa to miara, która ocenia wydajność wykorzystania zasobów radiowych. Można ją definiować jako liczbę transmitowanych bitów w jednostce czasu przy danej szerokości pasma. Dlatego celem samym w sobie jest maksymalizacja tej wartości, to jest przesyłanie jak największej liczby bitów przy zachowaniu tej samej szerokości pasma. Wartość tę można poprawić na wiele sposobów, np. poprzez opracowanie zaawansowanych metod alokacji zasobów radiowych, sterowanie mocą nadawczą urządzeń, wykorzystanie systemów wieloantenowych itp. Miara ta jest korzystna dla oceny działania systemu dynamicznego dostępu do widma, którego podstawowym założeniem jest jej poprawa względem metody statycznego przydział widma. Jednak najlepsze wyniki często wymagają dostępu do wielu dodatkowych informacji (tzw. informacji kontekstowych), takich jak pozycja użytkowników, historyczne statystyki obecności użytkowników, jakości ich usług itp. Przy stale rosnącym zapotrzebowaniu na usługi bezprzewodowe poprawa efektywności widmowej jest konieczna i nieunikniona. W konsekwencji informacje kontekstowe stają się bardzo ważnymi lub wręcz kluczowymi danymi potrzebnymi do osiągnięcia tego celu. Ten fakt z kolei stwarza potrzebę opracowania algorytmów i systemów, które gromadzą, przechowują, przetwarzają, aktualizują i, co najważniejsze, wykorzystują informacje kontekstowe. Jednocześnie kluczowym aspektem jest określenie, jakich dokładnie informacji kontekstowych wymaga dany algorytm lub system do działania. Ponadto w większości przypadków sam rodzaj informacji kontekstowych nie jest wystarczający, należy bowiem określić również format rejestrowanych informacji, ich dokładność, częstotliwość aktualizacji itp. W świetle tych faktów konieczne jest przyjrzenie się samej informacji kontekstowej, jej reprezentacji oraz sposobowi zarządzania nią [4].

Rysunek 1: Obszar analizy

W tym miejscu warto przyjrzeć się np. bazom środowiska radiowego, które z powodzeniem można wykorzystać w procesie alokacji zasobów radiowych. Tego typu bazy danych nie tylko przechowują informacje kontekstowe, ale także umożliwiają wnioskowanie na ich podstawie. Nie sposób nie wspomnieć również o istotnym obecnie trendzie, w którym dostęp radiowy jest zwirtualizowany. Koncepcja wirtualizacji, poprzez wyodrębnienie samego algorytmu zarządzania widmem, umożliwia implementacje tego algorytmu w postaci aplikacji. Podejście to skutkuje brakiem konieczności przeprojektowywania całego systemu w celu zmiany sposobu dostępu do widma.

Po połączeniu wszystkich przedstawionych powyżej aspektów zastosowanie dynamicznego podziału widma z wykorzystaniem informacji kontekstowych wydaje się obiecującym rozwiązaniem problemu niewystarczających zasobów radiowych i często nieefektywnego ich wykorzystania. Skłania to do postawienia następującej tezy badawczej: „Z pomocą informacji kontekstowej możliwe jest zaprojektowanie algorytmu zarządzania widmem, w wyniku którego zwiększa się efektywność widmowa systemu". W pracy zostanie przedstawione obszerne uzasadnienie postawionej tezy poprzez analizę teoretyczną i weryfikacje eksperymentalną samego systemu dostępu do widma. Strukturalny podział pracy pokazano na Rys. 1. Główny temat pracy jest koncepcyjnie podzielony na pięć części, w których każdej części poświęcony został osobny rozdział. Każdy z nich natomiast zawiera kilka podrozdziałów, w których opisane zostały przeprowadzone badania lub eksperymenty.

2. PRZETWARZANIE I PRZECHOWYWANIE INFORMACJI KONTEKSTOWEJ

Samo zagadnienie informacji kontekstowej to dość obszerny temat i dlatego, aby wprowadzić czytelnika, część druga poświęcona jest przede wszystkim opisowi informacji kontekstowych i aspektów z nimi związanych, w tym przykładu pozyskiwania, przechowywania i wykorzystywania takich informacji. W szczególności zaprezentowano tam pojęcia map środowiska radiowego oraz map usług radiowych, które posłużyły za podstawę działania systemu dynamicznego dostępu do widma rozważanego na początku części trzeciej. W tym przypadku mapy środowiska radiowego mogą być budowane na podstawie informacji o mocy odbieranej w paśmie telewizyjnym na analizowanym obszarze. Wykorzystanie zestawu pomiarowego umożliwiło wypełnienie omawianej mapy danymi rzeczywistymi. Następnie za pomocą interpolacji możliwe było uzupełnienie brakujących miejsc na mapie (tj. punktów bez powiązanych z nimi pomiarami) w analizowanym obszarze, tworząc kompletną mapę mocy odbieranego sygnału. Procedura ta pozwala zaproponować, zaprojektować i przetestować system dynamicznego dostępu do widma, który umożliwia wykorzystanie danego pasma do innych celów, tj. do nadawania danych nielicencjonowanych użytkowników, przy zachowaniu oryginalnej, trwającej już transmisji. To z kolei jest praktycznym przykładem wykorzystania informacji kontekstowych do poprawy efektywności widmowej na określonym obszarze.

3. PODEJŚCIE TEORETYCZNE W SYSTEMACH DSA

Część trzecia jest poświęcona teoretycznym rozważaniom na temat wykorzystania różnych rodzajów informacji kontekstowych do projektowania lub ulepszania działania dynamicznego systemu przydziału widma. Słowo „teoretyczny" zostało tutaj użyte w celu wskazania badań (symulacji i eksperymentów), które nie narzucają ograniczeń wynikających z wykorzystania określonej technologii transmisji, standardu sieci itp. Najgorszym przypadkiem dla efektywności widmowej jest sytuacja, gdy nikt nie używa wybranego fragmentu pasma, a jednocześnie ktoś inny chce uzyskać dostęp do widma. W tym przypadku często rozważa się metody współdzielenia widma oparte na wykrywaniu transmisji użytkowników pierwotnych. Detekcja ta pozwala zachować dotychczasowe (pierwotne) przeznaczenie danego fragmentu pasma, ale w przypadku chwilowego braku pierwotnych użytkowników – umożliwia wykorzystanie tego samego pasma do innych celów. Takie rozwiązanie prezentuje m.in. część trzecia, gdzie eksperyment z wykorzystaniem bazy danych środowiska radiowego z informacjami o mocy odbieranej w paśmie telewizyjnym pozwala użytkownikom nielicencjonowanym na

transmisję przy jednoczesnej ochronie licencjonowanych użytkowników [5]. Dzięki temu możliwe było prowadzenie transmisji wideo pomiędzy nielicencjonowanymi użytkownikami korzystającymi z pasma telewizyjnego w warunkach laboratoryjnych z zachowaniem możliwości odbioru telewizji naziemnej w tym samym miejscu. Warto w tym miejscu podkreślić, że jest to możliwe w wielu miejscach (zwłaszcza wewnątrz budynków) ze względu na często niskie wykorzystanie pasma telewizyjnego oraz często niski poziom sygnału odbieranego wewnątrz budynków (gdzie do prawidłowego odbioru telewizji niezbędna jest antena zewnętrzna). Kolejnym problemem jest rozdzielenie widma pomiędzy różnych użytkowników w taki sposób, aby odpowiadało to ich potrzebom przy jednoczesnej maksymalizacji efektywności widmowej. Przy założeniu dodatkowej możliwości sterowania mocą nadawczą użytkowników problem ten staje się bardzo złożony. W części trzeciej opisana została próba rozwiązania tego problemu, w której proponuje się dwustopniowe suboptymalne algorytmy do jednoczesnej alokacji zasobów i sterowania mocą. Dzięki wykorzystaniu informacji kontekstowych możliwe było zmaksymalizowanie średniej przepływności użytkowników na km^2, co zwiększa efektywność widmową na jednostkę powierzchni. Ponadto w tej części opisano algorytm, który optymalizuje wydajność widmową poprzez odpowiedni dobór kształtu fali dla użytkowników. Wybranie kształtu fali o mniejszym upływie mocy do sąsiednich kanałów pozwala na inny (często korzystniejszy) sposób przydzielania kanałów między użytkownikami. Algorytm wyboru kształtu fali użytkowników nielicencjonowanych poprawia wydajność widmową systemu poprzez zwiększoną elastyczność w przydzielaniu widma użytkownikom, co z kolei przekłada się na większą liczbę użytkowników korzystających z ta sama część widma. W tej części pokazano poprzez symulacje komputerowe i prosty eksperyment z wykorzystaniem rzeczywistego sprzętu, że bez wątpienia można poprawić efektywność widmową, zakładając obecność bazy informacji kontekstowych i stosując odpowiedni algorytm, który te informacje kontekstowe uwzględnia podczas pracy w systemie.

4. PODEJŚCIE PRAKTYCZNE W SYSTEMACH DSA

Część czwarta przedstawia praktyczne podejście do wykorzystania informacji kontekstowych w dynamicznym systemie dostępu do widma. Wspomniany aspekt „praktyczny" dotyczy przede wszystkim dodatkowych ograniczeń narzucanych przez już działającą sieć, gdzie nie możemy jej modyfikować, a dodatkowo musimy uwzględnić ograniczenia wprowadzane przez zastosowany standard, protokół lub system transmisji. W porównaniu z częścią trzecią, opisane tutaj eksperymenty są tutaj bliższe rzeczywistej implementacji w już funkcjonujących sieciach. Pierwszym opisanym eksperymentem są testy terenowe proponowanego dynamicznego systemu dostępu do widma, który umożliwia nowym nielicencjonowanym użytkownikom dostęp do zasobów radiowych przy jednoczesnej ochronie użytkowników głównych. Zaprojektowano kolejny system dostępu do widma, który tym razem w testach polowych w rzeczywistej sieci potwierdził możliwość korzystania z widma przez nielicencjonowanych użytkowników bez zauważalnego spadku jakość licencjonowanej transmisji. Ten eksperyment, przeprowadzony przy użyciu rzeczywistego sprzętu i komercyjnej radiowej sieci dostępu do Internetu, był kolejnym potwierdzeniem słuszności postawionej w pracy tezy. W tym przypadku głównymi użytkownikami są punkty dostępu do sieci radiowej zaimplementowanej w technologii WiMAX w kilku wsiach [3]. Drugim opisywanym eksperymentem jest symulacja komputerowa działającej w budynku sieci LTE, której zasoby są udostępniane użytkownikom spoza budynku [6]. Głównym ograniczeniem jest używany system transmisji LTE. Zaproponowano algorytm półtrwałego planowania, który korzystając z informacji kontekstowej o historii alokacji zasobów dla użytkowników, pozwala zaalokować zasoby niektórym użytkownikom na dłuższy czas (niż domyślnie). Powoduje to zmniejszenie ilości danych sterujących przesyłanych przez sieć poprzez zmniejszenie liczby decyzji planisty. Trzecim opisanym eksperymentem jest ponownie symulacja sieci bezprzewodowej; jednak w tym przypadku brane jest pod uwagę gęste środowisko radiowe i system wieloantenowy z bardzo dokładnym modelem kanału opartym na metodzie śledzenia promieni. Sposób przypisywania użytkowników do komórek jest modyfikowany na podstawie historycznych decyzji. Głównym ograniczeniem tego podejścia jest wykorzystanie sieci LTE-A, co znacznie ogranicza swobodę przypisywania użytkowników do komórek, poprzez konieczność sterowania odpowiednimi parametrami dostępnymi już w sieci. W tym przypadku wykorzystanie informacji kontekstowych poprawia przepływność wszystkich użytkowników sieci, przy niewielkim koszcie jakim jest większa liczba przełączeń między użytkowników między komórkami.

5. WYKORZYSTANIE W GĘSTYCH SIECIACH BEZPRZEWODOWYCH

W zakresie efektywności widmowej w sieciach radiowych nie sposób nie wspomnieć o obserwowanym od pewnego czasu trendzie (szczególnie w kontekście sieci 5G), który zakłada zwiększenie gęstości węzłów sieci i zmniejszenie rozmiaru komórek (zwiększenie ich liczby) [2]. Część piąta przedstawia rozważania dotyczące wykorzystania informacji kontekstowych w bardzo gęstych sieciach radiowych. Główną inspiracją do opisywanych prac jest obserwacja układu nerwowego człowieka (zwłaszcza mózgu), gdzie wykorzystywana jest bardzo prosta metoda transmisji, gdzie gęstość połączeń nerwowych jest ogromna. Oczywiście połączenia neuronowe są znacznie bardziej podobne do przewodowego systemu transmisji, ale wiele obserwacji z układu nerwowego można nadal wykorzystywać w bezprzewodowych systemach transmisji. W szczególności wyróżniono dwa dodatkowe typy komórek w ludzkim układzie nerwowym, tj. astrocyty i mikroglej. Astrocyt działa jako rodzaj łącznika między dwoma niezależnymi układami (układem krążenia i układem nerwowym) i możliwość wpływu jednego układu na drugi drugiego. Z drugiej strony mikroglej jest niezbędną komórką układu odpornościowego, która początkowo jest w stanie spoczynku. Jednak w przypadku wykrycia anomalii układu nerwowego, zamienia się ona w aktywną formę i podejmuje próbę „naprawy" tej anomalii. Inspiracja płynąca z opisanych komórek pozwala przede wszystkim na poprawę niezawodności sieci bezprzewodowej. W przypadku tej tezy informacja kontekstowa jest dwojaka. Węzeł inspirowany mikroglejem wykrywa problemy z transmisją (np. utratę transmisji sąsiedniego węzła). Natomiast w przypadku węzła inspirowanego astrocytami informacje zewnętrzne (spoza sieci) mogą pochodzić np. z czujników i mieć wpływ na działanie sieci. Efektywność widmowa jest w tym przypadku ściśle związana z niezawodnością

transmisji. Warto zauważyć, że widmo jest efektywnie wykorzystywane, gdy za jego pomocą przesyłanych jest dużo informacji. Jednak dane przekazane przez sieć muszą być prawidłowe.

6. WIRTUALIZACJA DOSTĘPU DO WIDMA

Kolejnym istotnym trendem w telekomunikacji jest obecnie (obserwowana od pewnego czasu) wirtualizacja dostępu radiowego, gdzie poszczególne funkcje sieci pełni oprogramowanie. W ostatniej, ale równie ważnej szóstej cześci, opisano kilka podejść do zarządzania widmem, lecz reprezentowanych w formie aplikacji do zarządzania widmem. Najpierw opisana jest aplikacja zarządzająca widmem używana w połączeniu z platformą Open Air Interface, która na podstawie wybranego zestawu reguł i zasad dobiera optymalny (według kryteriów określonych przez operatora) zestaw parametrów transmisji dla każdego użytkownika, a w szczególności zakres częstotliwości. Opisano autorski algorytm podejmowania decyzji wraz z implementacją na rzeczywistych urządzeniach (telefony komórkowe). W tym przypadku nie bierzemy pod uwagę bezpośrednio informacji kontekstowych - zaznaczmy jednak, że wybrany zestaw zasad ma zasadniczy wpływ na działanie analizowanego systemu, tj. jego wydajność widmową. Z kolei nawet w najprostszym scenariuszu zestawy zasad wybierane są na podstawie pewnych informacji kontekstowych. W najbardziej zautomatyzowanej formie na podstawie zebranych danych kontekstowych, algorytmy uczenia maszynowego dobierają odpowiednie zestawy reguł. Jest to przykład pośredniego wykorzystania informacji kontekstowych na wyższym poziomie zarządzania widmem. Następnie, w części szóstej, przedstawiono nowe podejście do współdzielenia widma, w którym niektóre pasma widma są dostępne bezpłatnie dla wszystkich użytkowników, a reszta widma jest dostępna za opłatą. Takie podejście zostało opisane z perspektywy użytkownika i właściciela takiej sieci („właściciela" widma). Podobnie jak w przypadku opisanej wcześniej sytuacji, nie bierzemy pod uwagę bezpośrednio informacji kontekstowych, ale zbiory reguł. To z kolei implikuje zmianę efektywności widmowej analizowanego systemu. Ostatnią pracą opisaną w tej części jest zastosowanie informacji kontekstowych w otwartej sieci dostępu radiowego [1]. Jest to doskonały przykład, w którym cały system jest zwirtualizowany i możliwe jest dołączenie do takiego systemu aplikacji (tzw. xApp), które implementują określone funkcjonalności. Opisany został przykład aplikacji, która steruje ruchem użytkowników poprzez zmianę ich przypisania do komórek, zmianę przydziału zasobów radiowych i jednocześnie wybieranie pasma wykorzystywanego przez poszczególne komórki. Przedstawione w tej części koncepcje pokazują, że zmieniając sposób przypisywania pasm do stacji bazowych, przypisywania użytkowników do komórek, czy przydzielania użytkownikom zasobów radiowych, można poprawić wydajność widmową całego systemu. Warto jednak zauważyć, że musi to być to aktualny cel zdefiniowany przez zbiór zasad.

7. PODSUMOWANIE

W tej pracy zostało pokazane z wielu perspektyw i na wiele sposobów, że istotnie możliwa jest poprawa efektywności widmowej w sieci bezprzewodowej poprzez wykorzystanie informacji kontekstowych, co potwierdza tezę postawioną w tej pracy. Dobór metod uzyskiwania, przechowywania i wykorzystywania informacji kontekstowej, jak i wybór jaka informacja kontekstowa jest potrzebna, powinien być rozważany każdorazowo dla analizowanego systemu.

LITERATURA

[1] Gavrilovska Liljana, Rakovic Valentin, Denkovski Daniel. 2020. "From cloud RAN to open RAN". Wireless Personal Communications, 113 : 1523–1539.

[2] Kamel Mahmoud, Hamouda Walaa, Youssef Amr. 2016. "Ultra-dense networks: A survey". IEEE Communications Surveys Tutorials, 18 (4) : 2522–2545.

[3] Kliks Adrian, Kryszkiewicz Paweł, Kułacz Łukasz, Kowalik Karol. 2019. "Database supported flexible spectrum access - field trials in commercial networks". IEEE International Conference on Communications Workshops (ICC Workshops) : 1–6.

[4] Schilit Bill, Hilbert David, Trevor Jonathan. 2002. "Context-aware communication". IEEE Wireless Communications. 9 (5) : 46–54.

[5] Sohul Munawwar, Yao Miao, Yang Taeyoung, Reed Jeffrey. 2015. "Spectrum access system for the Citizen Broadband Radio Service". IEEE Communications Magazine, 53 : 18–25.

[6] Ye Yun, Wu Dalei, Shu Zhihui, Qian Yi. 2016. "Overview of LTE Spectrum Sharing Technologies" IEEE Access, 4 : 8105–8115.

MIKROFALOWE SPRZĘGACZE KIERUNKOWE O POLEPSZONYCH PARAMETRACH I ICH ZASTOSOWANIE W MONOLITYCZNYCH MIKROFALOWYCH UKŁADACH SCALONYCH

MICROWAVE DIRECTIONAL COUPLERS HAVING IMPROVED PERFORMANCE AND THEIR APPLICATION IN MONOLITHIC MICROWAVE INTEGRATED CIRCUITS (MMIC)

Robert Smolarz

Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

DOI: 10.15199/59.2023.4.31

Streszczenie: **Niniejszy referat opisuje wybrane zagadnienia dotyczące projektowania mikrofalowych sprzęgaczy kierunkowych o polepszonych parametrach i ich zastosowanie w monolitycznych mikrofalowych układach scalonych (ang. Monolithic Microwave Integrated Circuit). W pierwszej części publikacji, przedstawiono opracowaną metodę kompensacji, która pozwala na zmniejszenie strat odbiciowych oraz zwiększenie izolacji sprzęgaczy zbliżeniowych.Kolejnym z poruszonych zagadnień jest projektowanie zminiaturyzowanych sprzęgaczy gałęziowych pobudzanych różnicowo. Ostatnia kwestia jaka została poruszona w ramach referatu, związana jest z integracją opracowanych sprzęgaczy o polepszonych parametrach w układach scalonych, które mogą znaleźć swoje zastosowanie w technice sensorowej, radiolokacyjnej i radiokomunikacyjnej. Wspomniana tematyka została szerzonej opisana w rozprawie doktorskiej autora pt. „*Microwave directional couplers' design with the use of planar quasi-TEM transmission line sections. Analysis, experimental investigation and applications*”. Praca ta opisuje szereg zaproponowanych, nowatorskich rozwiązań projektowych dotyczących sprzęgaczy kierunkowych projektowych w technice PCB oraz MMIC.**

Abstract: **The paper describes selected design methodologies utilized in microwave directional couplers with improved parameters and their application in MMIC circuits. In the first of the paper, the developed capacitive compensation method is presented, which allows to reduce return losses and increase the isolation of coupled-line sections. Second section is focused on the design of miniaturized differentially excited branch-line couplers. The last section is related to the integration of the developed couplers in integrated circuits that can be used in sensor, radar and radiocommunication devices. The discussed issues have been extensively described in the author's doctoral dissertation entitled „Microwave directional couplers' design with the use of planar quasi-TEM transmission line sections. Analysis, experimental investigation and applications”, which work describes a number of innovative design solutions for design directional couplers in PCB and MMIC technology.**

Słowa kluczowe: **sprzęgacze kierunkowe, MMIC, miniaturyzacja, kompensacja pojemnościowa, sekcja sprzężona, sprzęgacze pobudzane różnicowo.**

Keywords: **directional couplers, MMIC, miniaturized directional couplers, capacitive compensation method, coupled-line section, differentially-fed directional couplers.**

1. WSTĘP

Sprzęgacze kierunkowe należą do jednych z bardziej popularnych komponentów pasywnych stosowanych w technice mikrofalowej. Komponenty te przede wszystkim pozwalają na uzyskiwanie odpowiedniego, wymaganego podziału mocy mikrofalowej oraz przesunięć fazowych między wrotami. Jako komponenty składowe bardziej złożonych obwodów pracujących w pasmach radiowych, znajdują swoje zastosowanie w nowoczesnych rozwiązaniach radiokomunikacyjnych, radiolokacyjnych czy pomiarowych. Ze względu na coraz wyższe wymagania stawiane przed urządzeniami radiowymi, naturalnym staje się fakt polepszenia właściwości samych elementów składowych jakimi są między innymi omawiane sprzęgacze kierunkowe.

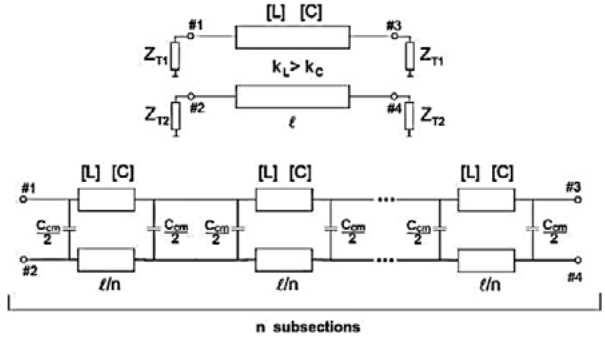
Z punktu widzenia wykorzystywanych topologii, sprzęgacze kierunkowe można podzielić na te oparte na idei sekcji linii sprzężonych (ang. *coupled-lines section*), które są również znane jako sprzęgacze zbliżeniowe oraz tzw. bezpośrednio połączonych linii transmisyjnych (ang. *directly connected-lines)* [1]. Do tej drugiej grupy zalicza się między innymi sprzęgacze gałęziowe i pierścieniowe.

Sprzęgacze kierunkowe zaprojektowane w oparciu o pojedynczą sekcję sprzężoną charakteryzują się względnie szerokim pasmem pracy, sięgającym nawet jednej oktawy częstotliwości. Jeśli chodzi o wady tego rozwiązania to przede wszystkim należy tutaj wymienić zmniejszanie izolacji oraz zwiększanie strat odbiciowych wraz ze wzrostem częstotliwości, co wiąże się też z nierównoważnością między sprzężniami: indukcyjnym i pojemnościowym ($k_L \neq k_C$). W celu zminimalizowania tego efektu, który jest spotykany w rzeczywistych strukturach sprzężonych, literatura podaje kilka rozwiązań, takich jak: zastosowanie odpowiedniego uwarstwienia struktury [2], użycie dodatkowych elementów skupionych i quasi-skupionych [3], periodyczną modyfikację sekcji sprzężonej [4] czy zastosowanie linii opóźniającej [5].

Kolejną kwestią jest miniaturyzacja układowa, która jest często porządaną własciwością w przypadku techniki MMIC (ang. *Monolithic Microwave Integrated Circuit).* Rozpatrując sprzęgacze gałęziowe, literatura podaje kilka rozwiązań, które są dedykowane technologiom monolitycznym. Należy natomiast wspomnieć, że większość opracowanych sprzęgaczy kierunkowych tego typu to rozwiązania pobudzane w sposób nieróżnicowy.

W niniejszym referacie przedstawiono dwa wybrane zagadnienia dotyczące projektowania sprzęgaczy kierunkowych w technice MMIC, które zostały zaproponowane w publikacjach [6] oraz [7] i które były również obiektem rozpatrywań rozprawy doktorskiej autora. Pierwsze z nich tyczy się możliwości poprawy parametrów jakimi są izolacja oraz straty odbiciowe sprzęgaczy kierunkowych opartych na sekcjach sprzężonych za pomocą opracowanej metody kompensacji pojemnościowej. Drugi z poruszonych aspektów dotyczy miniaturyzacji gałęziowych sprzęgaczy kierunkowych

pobudzanych w sposób różnicowy. Opracowane rozwiązania zostały zintegrowane w bardziej złożonych obwodach co też zostało pokrótce omówione w niniejszej publikacji.



Rys. 1. Schemat ideowy przedstawiający sekcję sprzężoną złożoną z dwóch linii transmisyjnych oraz układ przedstawiaający tę samą sekcję podzieloną na n-subsekcji, który został dodatkowo skompensowany pojemnościami C_{cm} [6].

2. KOMPENSACJA POJEMNOŚCIOWA W SCALONYCH SPRZĘGACZACH KIERUNKOWYCH NA PRZYKŁADZIE TRÓJPASKOWEJ SEKCJI SPRZĘŻONEJ

Rysunek 1 przedstawia uproszczony schemat dwóch sprzężonych linii transmisyjnych, który można opisać za pomocą macierzy indukcyjnej [L] oraz pojemnościowej [C]. Rozpatrując przypadek, w którym sprzężenie indukcyjne k_L jest większe od pojemnościowego k_C, układ ten może zostać skompensowany za pomocą metody pojemnościowej [6]. W tym celu, prosta sekcja sprzężona może zostać zastąpiona za pomocą n-subsekcji sprzężonych, posiadających dodatkowe jednostkowe pojemności kompensujące C_{cm}, które znajdują się między liniami sprzężonymi (układ ten został zaprezentowany również na rys. 1). Aby zrozumieć ideę zaproponowanej kompensacji pojemnościowej, w pierwszej kolejności należy rozpatrzeć przypadek idealnej sekcji sprzężonej dla której występuje następująca zależność:

$$k_{required} = k_L = k_C, \tag{1}$$

W powyższym równaniu, $k_{required}$ oznacza wymagane sprzężenie sekcji sprzężonej. Sprzężenia $k_{required}$, k_L i k_C mogą zostać obliczone z następującej zależności uwzględniającej pojemność kompensującą C_d [6]:

$$k_{required}{}^2 = k_L^2 = k_C^2 = \frac{(C_{12}+C_d)^2}{(C_{11}+C_d)(C_{22}+C_d)}. \tag{2}$$

Pojemności C_{11}, C_{22}, C_{12} są współczynnikami macierzy pojemnościowej rzeczywistej, nieskompensowanej sekcji sprzężonej. Z powyższego równania, można wyznaczyć pojemność kompensującą C_d, a następnie oszacować wartości impedancji sprzężonych linii transmisyjnych, zgodnie z zależnościami [6]:

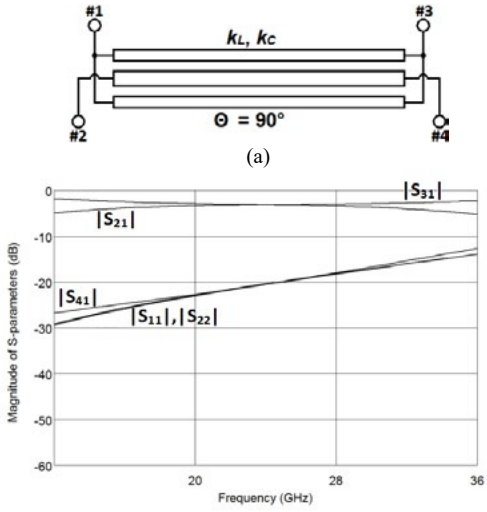
$$Z_{L1} = \sqrt{\frac{L_{11}}{C_{11}+C_d}}, \quad Z_{L2} = \sqrt{\frac{L_{22}}{C_{22}+C_d}}, \tag{3}$$

gdzie Z_{L1}, Z_{L2} oznaczają impedancje linii transmisyjnych, a indukcyjności L_{11} i L_{22} są współczynnikami macierzy indukcyjnej [L]. Pojemność kompensującą Cd realizuje się za pomocą pojemności skupionych oznaczonych jako C_{cm}. Wartość C_{cm} można oszacować w odniesieniu do długości fizycznej sekcji sprzężonej l oraz n-tej ilości quasi-skupionej pojemności kompensującej C_d:

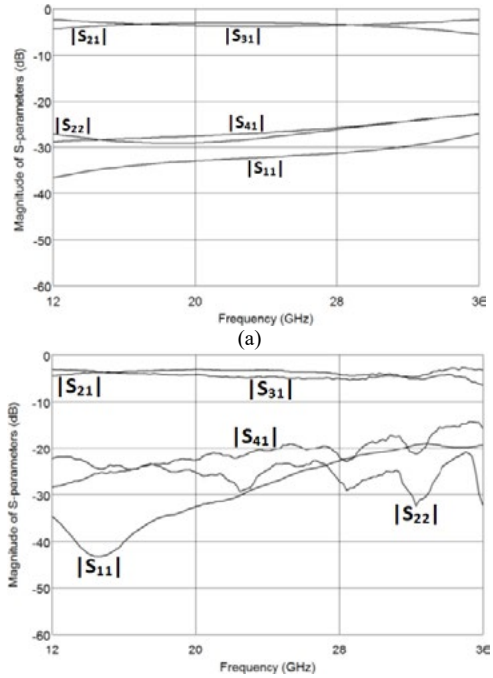
$$C_{cm} = \frac{C_d}{n} l \tag{4}$$

W porównaniu do znanych rozwiązań opierających się na kompensacji pojemnościowej i indukcyjnej, zaproponowane rozwiązanie pozwala na znaczące zredukowanie ilości elementów kompensujących. Ta cecha została wykorzystana w sprzęgaczach kierunkowych zaprojektowanych w technice monolitycznej, gdzie ilość elementów ma znaczący wpływ na miniaturyzację oraz stratność całego układu.

Przedstawiona metoda została użyta do zaprojektowania trzech scalonych sprzęgaczy zbliżeniowych, wykonanych w trzech topologiach: w strukturze trójpaskowej, w topologii Langego oraz w strukturze lewoskrętnych linii transmisyjnych [6]. Wszystkie wymienione sprzęgacze zostały zaprojektowane w procesie UMS PH25 bazującym na arsenku galu. Na potrzeby tej publikacji, do prezentecji rezultatów, wybrano jedynie pierwszą topologię.



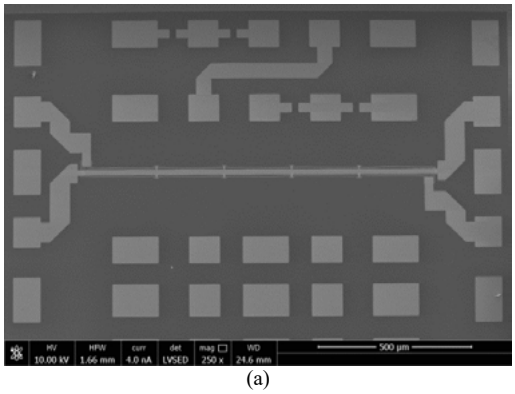
Rys. 2. Schemat koncepcyjny zbliżeniowego sprzęgacza kierunkowego wykonanego w strukturze trzech sprzężonych linii transmisyjnych (a) oraz jego odpowiedź częstotliwościowa (b) [6].



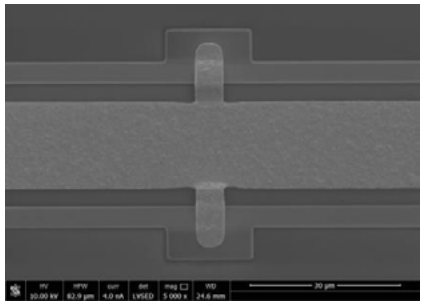
Rys. 3. Wyniki pełnofalowych symulacji elektromagnetycznych (a) oraz pomiarów (b) monolitycznego sprzęgacza kierunkowego zaprojektowanego w strukturze trzech sprzężonych linii transmisyjnych [6].

Koncepcyjny schemat sprzęgacza kierunkowego opartego o sekcję trzech sprzężonych linii transmisyjnych został zaprezentowany na rys. 2 (a). Zwiększenie ilości sprzężonych linii transmisyjnych wpływa bezpośrednio na maksymalne, możliwe do zrealizowania sprzężenie. W przypadku procesów monolitycznych (scalonych), uzyskanie silnego sprzężenia w strukturze jest często bardzo trudne do uzyskania ze względu na ograniczenia technologiczne (m.in. minimalną odległość między liniami, czy maksymalną ilość warstw metalizacji). Toteż, rozpatrywana struktura została wykorzystana w celach weryfikacyjnych w technologii monolitycznej.

Rysunek 2 (b) przedstawia odpowiedź częstotliwością nieskompensowanej struktury trójpaskowej. Jak widać, izolacja opisana parametrem rozproszenia S_{41} zmniejsza się wraz ze wzrostem częstotliwości. Natomiast straty odbiciowe opisywane za pomocą parametrów S_{11} i S_{22} - wzrastają. Oczywistym jest, że takie efekty są niepożądane i znacząco mogą wpływać na parametry i funkcjonowanie układów, urządzeń w których sprzęgacze mogą zostać użyte. W celu polepszenia ww. parametrów, w trakcie projektowania rozpatrywanej struktury, zastosowano zaproponowaną metodę kompensacji pojemnościowej. Wyniki symulacji pełnofalowych oraz pomiarów zostały zaprezentowane kolejno na rys. 3(a) oraz rys. 3(b). Jak można zauważyć, zastosowanie metody kompensacyjnej znacząco poprawia właściwości sprzęgacza, szczególnie w zakresie wyższych częstotliwości. Rysunek 4(a) przedstawia zdjęcie wykonanej struktury, natomiast 4(b) ukazuje realizację quasi-skupionej pojemności kompensującej C_{cm}, która została zrealizowana za pomocą dwunastu mostków powietrznych.

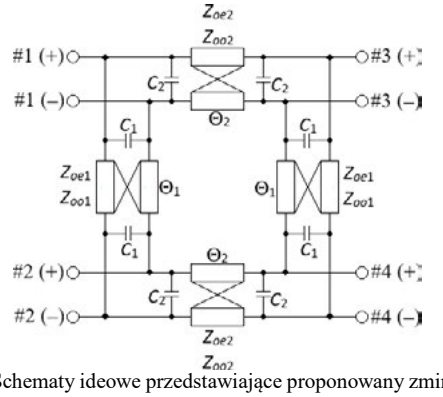


(a)



(b)

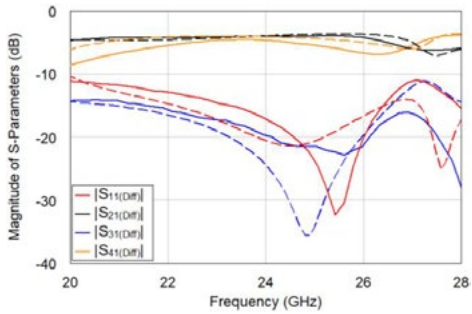
Rys. 4. Zdjęcia prezentujące wykonany sprzęgacz kierunkowy (a) oraz zbliżenie na jeden z mostków powietrznych realizujących pojemność kompensującą (b) [6].



Rys. 5. Schematy ideowe przedstawiające proponowany zminiaturyzowany sprzęgacz pobudzany różnicowo [7].

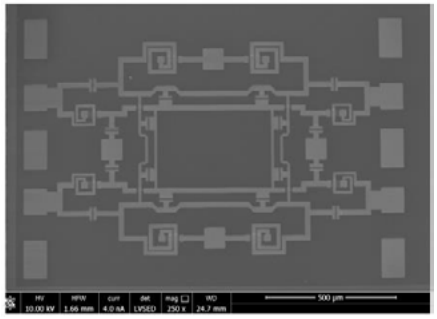
3. ZMINIATURYZOWANY, MONOLITYCZNY GAŁĘZIOWY SPRZĘGACZ KIERUNKOWY POBUDZANY RÓŻNICOWO

Klasyczny sprzęgacz gałęziowy jest strukturą pobudzaną w sposób nieróżnicowy i składa się z czterech połączonych linii transmisyjnych posiadających impedancję charakterystyczną równą 35.35 Ω i 50 Ω. Rozwiązanie zaproponowane w [7] jest układem pobudzanym różnicowo, gdzie sam sprzęgacz jest złożony z czterech sekcji sprzężonych. Dodatkowo, w celu miniaturyzacji układowej, zastosowano metodę zaprezentowaną w [8], gdzie autorzy użyli dodatkowych pojemności skupionych w celu efektywnego skrócenia długości elektrycznej linii w sprzęgaczu pobudzanym nieróżnicowo. Schemat zaproponowanego zminiaturyzowanego sprzęgacza pobudzanego w sposób różnicowy został zaprezentowany na Rys. 5.



Rys. 6. Odpowiedź częstotliwościowa zminiaturyzowanego, scalonego sprzęgacza różnicowego zaobserwowana w wyniku symulacji (linie przerywane) oraz pomiarów (linie ciągłe) [7].

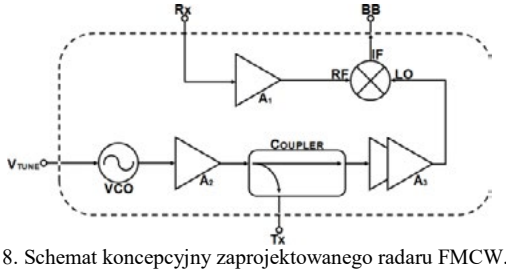
Charakterystyki przedstawione na Rysunku 6 ilustrują symulowaną elektromagnetycznie oraz zmierzoną odpowiedź częstotliwością sprzęgacza. Na podstawie uzyskanych wyników można stwierdzić, że izolacja mierzona na częstotliwości środkowej nie jest mniejsza niż 20 dB, a dopasowanie nie jest gorsze niż 17.7 dB. Są to wyniki w pełni zadowalające. Dodatkowo symulacje oraz pomiary w znaczącym stopniu pokrywają się. Zdjęcie wykonanego układu zostało zaprezentowane na Rys. 7. Należy również wspomnieć o samym stopniu zminiaturyzowania, który w porównaniu do klasycznej struktury składającej się z linii ćwierćfalowych wynosi ponad 50 %.



Rys. 7. Zdjęcie wykonane mikroskopem TEM przedstawiające wykonany zminiaturyzowany sprzęgacz różnicowy [7].

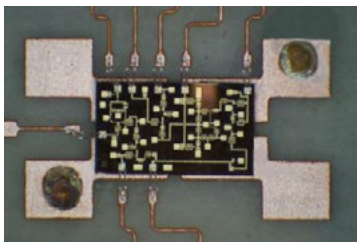
4. ZASTOSOWANIE NOWATORSKICH MONOLITYCZNYCH SPRZĘGACZY KIERUNKOWYCH W MIKROFALOWYCH UKŁADACH SCALONYCH

Sprzęgacze omówione we wcześniejszych rozdzialach zostały wykorzystane do zaprojektowania kolejno: monolitycznego radardu FMCW oraz wzmacniacza pobudzanego różnicowo. Oba układy zostały zaprojektowane dla pasma operacyjnego 22-25 GHz w technologii scalonej UMS PH25.



Rys. 8. Schemat koncepcyjny zaprojektowanego radaru FMCW.

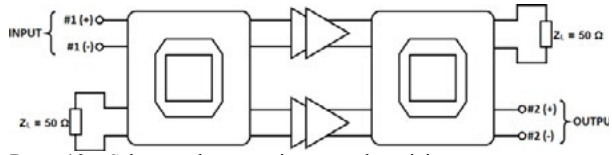
Rysunek 8 przedstawia koncepcję zaprojektowanego radaru FMCW pracującego na 24 GHz oraz charakteryzującego się maksymalną mocą wyjściową równą 0 dBm. Obwód składa się z generatora przestrajanego napięciem (VCO), mieszacza diodowego, wzmacniaczy $A_{1\text{-}3}$ oraz sprzęgacza zbliżeniowego oparte o ideę sprzężonych trzech linii transmisyjnych o którym była mowa w rozdziale 2. Rysunek 9 przedstawia wykonany układ scalony, który został przymocowany do płytki testowej.



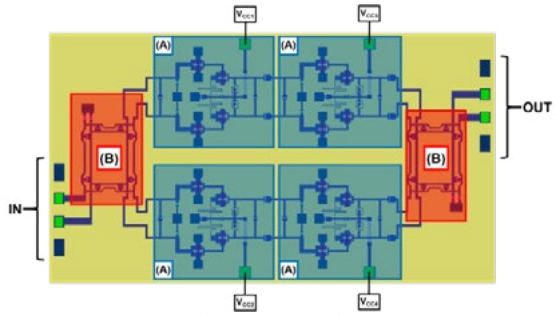
Rys. 9. Zdjęcie wykonanego radaru scalonego FMCW zamontowanego na płytce testowej.

Zminiaturyzowany sprzęgacz gałęziowy pobudzany różnicowo przedstawiony w rozdziale 3, został natomiast wykorzystany do zaprojektowania zrównoważonego wzmacniacza różnicowego. Rysunek 10 przedstawia schemat koncepcyjny takiego wzmacniacza, natomiast na rys. 11, zaprezentowano layout gotowego układu.

Zarówno w przypadku zaprojektowanego i wykonanego radardu FMCW, jak i różnicowego wzmacniacza zrównoważonego, uzyskane wyniki symulacji oraz pomiary były w pełni satysfakcjonujące (m.in. uzyskano maksymalne wzmocnienie mocy na poziomie 21 dB w przypadku wzmacniacza). Pełne wyniki symulacji oraz pomiary zostały przedstawione w ogólnodostępnej rozprawie doktorskiej autora.



Rys. 10. Schemat koncepcyjny przedstawiający proponowany różnicowy wzmacniacz zrównoważony.



Rys. 11. Layout omawianego wzmacniacza zrównoważonego na którym zaznaczono użyte bloki wzmacniające (A) oraz zminiaturyzowane sprzęgacze gałęziowe (B).

5. PODSUMOWANIE

W ramach niniejszego referetatu, przedstawiono dwa wybrane zagadnienia zaprezentowane w rozprawie doktorskiej autora [6], dotyczące poprawy parametrów elektrycznych monolitycznych sprzęgaczy zbliżeniowych oraz miniaturyzacji sprzęgaczy gałęziowych pobudzanych różnicowo [7]. Przedstawiono również implementację wyżej wymienionych układów w dwóch wybranych obwodach tzn. w radarze FMCW oraz wzmacniaczu zrównoważonym.

LITERATURA

[1] D. M. Pozar, Microwave engineering. John wiley & sons, 2009.

[2] M. Horno and F. Medina, “Multilayer planar structures for high-directivity directional coupler design,” *IEEE transactions on microwave theory and techniques*, vol. 34, no. 12, pp. 1442–1449, 1986.

[3] S.-F. Chang, J.-L. Chen, Y.-H. Jeng and C.-T. Wu, “New high-directivity coupler design with coupled spurlines,” *IEEE microwave and wireless components letters*, vol. 14, no. 2, pp. 65–67, 2004.

[4] S. Gruszczynski and K. Wincza, “Generalized methods for the design of quasi-ideal symmetric and asymmetric coupled-line sections and directional coplers,” *IEEE tranactions on microwave theory and techniques,* vol. 59, no. 7, pp. 1709–1718, 2011.

[5] H. Chun, J.-Y. Moon, S.-W. Yun and J.-K. Rhee, “Microstrip line directional couplers with high directivity,” *Electronics Letters,* vol. 40, no. 5, pp. 317–318, 2004.

[6] S. Gruszczynski, R. Smolarz, and K. Wincza, "Realization of high-performance broadband quadrature directional couplers in UMS PH25 technology," *Electronics*, vol. 8, no. 12, pp. 1520-, December 2019.

[7] S. Gruszczynski, R. Smolarz, C. Wu, and K. Wincza, "Monolithic Miniaturized Differentially-Fed Branch-Line Directional Coupler in GaAs Monolithic Technology," *Electronics*, vol. 9, no. 3, pp. 446-, March 2020.

[8] M. Gillick, et. all, “CPW two-stage balanced monolithic 15 GHz amplifier using reduced-size branchline couplers,” *in Proc of 23rd EuMC Conference*, 1993, pp. 764–766.

ZASTOSOWANIE ELEKTROD GRAFENOWYCH W PRZEŁĄCZNIKACH NA PASMO FAL MILIMETROWYCH WYKONANYCH W STRUKTURACH AlGaN/GaN

APPLICATION OF GRAPHENE ELECTRODES FOR MILLIMETER WAVE SWITCHES IN AlGaN/GaN STRUCTURES

Jakub Sobolewski[1]

[1] Politechnika Warszawska, Instytut Radioelektroniki i Technik Multimedialnych, Warszawa, jakub.sobolewski@pw.edu.pl

DOI: 10.15199/59.2023.4.32

Streszczenie: **W artykule przedstawiono wyniki badań nad możliwością zastosowania elektrod sterujących wykonanych z grafenu w przełącznikach w.cz. na zakres fal milimetrowych. W badaniach wykorzystano struktury testowe przełączników zintegrowanych z linią koplanarną wykonanych w strukturze półprzewodnikowej AlGaN/GaN w topologii szeregowej i bocznikowej. Na podstawie uzyskanych wyników wykazano możliwości wykorzystania zjawiska zmiany przewodności grafenu pod wpływem napięcia do ograniczenia strat w przełączniku. Wskazano także na możliwość uzyskania poprawy izolacji przy wykorzystaniu elektrod grafenowych. Konieczne jest jednak udoskonalenie procesów technologicznych dla uzyskania wysokiej jakości warstw grafenu.**

Abstract: **This paper presents research on application of graphene control electrodes for millimeter wave switches. Series and shunt topology switches integrated with coplanar waveguide in AlGaN/GaN semiconductor structure were examined. Presented results show that it is possible to utilize voltage-controlled conductivity of graphene to reduce losses through gate coupling. They also indicate superior performance of shunt topology and show that application of graphene gate can improve switch isolation. However, to achieve good performance, technological processes have to be improved to increase the quality of graphene layers.**

Słowa kluczowe: **AlGaN/GaN, fale milimetrowe, GaN, grafen, HEMT, przełącznik w.cz.,**

Keywords: **AlGaN/GaN, GaN, graphene, HEMT, millimeter wave, RF switch**

1. WSTĘP

Jedną z istotnych dróg rozwoju współczesnych systemów radiowych jest możliwość ich dynamicznej rekonfiguracji w celu szybkiego adaptowania się do zmiennych warunków pracy. Podstawowymi komponentami umożliwiającymi szybkie zmiany konfiguracji są przełączniki mikrofalowe. Wśród ich zastosowań wymienić można m.in. anteny inteligentne z przełączanymi wiązkami i polaryzacjami, szyki antenowe z modulacją czasową, wielopasmowe układy nadawczo-odbiorcze oraz układy z dupleksem w dziedzinie czasu [1].

Wraz z postępującym zagospodarowaniem zakresu fal milimetrowych wzrasta zapotrzebowanie na pracujące w tym zakresie przełączniki. W literaturze znaleźć można wiele rozwiązań tego typu przełączników w różnych technologiach. Dominujące jest jednak zastosowanie tranzystorów wykonanych z półprzewodników z grup III-V, a coraz częściej także krzemu [2]. Rosnącym zainteresowaniem cieszą się układy na bazie azotku galu (GaN) ze względu na możliwość uzyskania wysokich częstotliwości pracy, krótkich czasów przełączania oraz niskiej rezystancji, a jednocześnie pracy w wysokiej temperaturze i z sygnałami o znacznej mocy [3].

Podobnie jak w wielu innych obszarach elektroniki, także w badaniach nad przełącznikami znaczną uwagę poświęca się obecnie zastosowaniu tzw. materiałów dwuwymiarowych takich jak grafen. Materiały te tworzą bardzo cienkie warstwy, o grubościach rzędu pojedynczych atomów, które wykazują właściwości odmienne od tej samej substancji w innych postaciach. Grafen, pomimo skrajnie niskiej grubości zachowuje bardzo wysoką przewodność elektryczną i cieplną, a ponadto umożliwia zmianę rezystancji pod wpływem napięcia. W większości publikacji poświęconych wykorzystaniu grafenu w układach przełączających wykorzystuje się ostatnią z wymienionych właściwości w celu zastąpienia grafenem elementu aktywnego (tranzystora lub diody) [4, 5]. Innym spotykanym rozwiązaniem jest wykorzystanie membran grafenowych w układach typu MEMS [6].

W niniejszej pracy zbadano odmienną koncepcję wykorzystania grafenu w konstrukcji przełącznika - jako elektrody sterującej (bramki) elementu aktywnego w postaci tranzystora. W takiej konfiguracji niska grubość grafenu może przyczynić się do redukcji elementów pasożytniczych w elemencie przełączającym negatywnie wpływających na wysokoczęstotliwościowe parametry przełącznika. Jednocześnie grafen zachowuje wysoką przewodność niezbędną do szybkiego przełączania tranzystora. Ponadto możliwość zmiany przewodności grafenu pod wpływem przyłożonego napięcia może stanowić dodatkowy czynnik redukujący wpływ obecności bramki tranzystora na straty i izolację przełącznika.

W niniejszej pracy zbadano możliwość zastosowania bramki grafenowej w konstrukcji przełączników SPST na zakres fal milimetrowych wykonanych w heterostrukturach AlGaN/GaN. W tym celu zaprojektowano i zbadano struktury testowe przełączników o topologii bocznikowej z rozproszonym elementem aktywnym wyposażonych w elektrody bramki wykonane grafenu oraz

analogiczne struktury z bramką wykonaną z metalu, co umożliwia porównanie ich parametrów. Ponadto na podstawie badań struktur testowych z przełącznikami o topologii szeregowej zbadano możliwość ograniczenia wpływu bramki na parametry przełącznika poprzez odpowiednią polaryzację elektrody grafenowej.

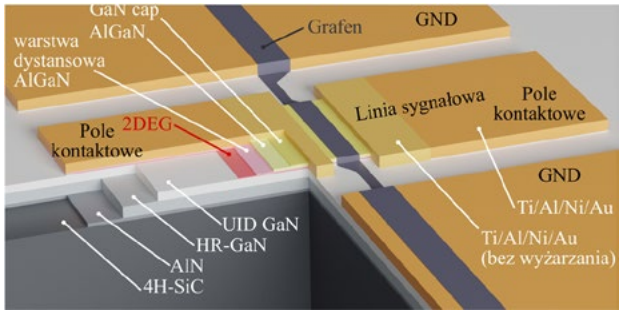
2. KONSTRUKCJA PRZEŁĄCZNIKÓW

Wszystkie opisywane struktury testowe zostały wykonane w strukturach półprzewodnikowych AlGaN/GaN opisanych szczegółowo w [7]. W strukturach tego typu na styku dwóch rodzajów półprzewodnika powstaje heterozłącze, w którym na skutek różnic szerokości pasma zabronionego występuje wąska studnia potencjału. W połączeniu z polaryzacją piezoelektryczną złącza powoduje ona wytworzenie warstwy o bardzo wysokiej koncentracji i ruchliwości elektronów nazywanej dwuwymiarowym gazem elektronowym (2-dimensional electron gas, 2DEG) stanowiącej kanał tranzystora. Umieszczenie na warstwie AlGaN elektrod wykonanych z metalu bądź grafenu powoduje utworzenie złącza Schottky'ego, którego polaryzacja w kierunku zaporowym umożliwia obniżenie koncentracji elektronów w warstwie 2DEG i w konsekwencji zwiększenie rezystancji kanału (wyłączenie tranzystora).

Aby zapewnić niskie straty i możliwość łatwego łączenia zaprojektowanych przełączników z innymi elementami toru sygnałowego oraz aparaturą pomiarową zaproponowano konstrukcję przełączników z elementami aktywnymi zintegrowanymi z koplanarną linią transmisyjną (CPW).

2.1. Przełącznik o topologii szeregowej

W celu zbadania możliwości zmniejszenia wpływu obecności bramki na parametry przełącznika poprzez wykorzystane właściwości zmiany przewodności grafenu pod wpływem napięcia wykorzystano przełącznik o topologii szeregowej (rys. 1) opisany w [7]. W tego typu konstrukcji tranzystor włączony jest w szereg z centralnym przewodnikiem CPW umożliwiając jego przerywanie. Bramka, umieszczona w poprzek linii i dołączona do bocznych przewodników (masy) CPW może tu powodować znaczne zaburzenia impedancji linii, a także niepożądane sprzężenie pojemnościowe w stanie wyłączonym przełącznika. Obniżenie przewodności bramki w stanie wyłączonym może mieć zatem szczególnie istotne znaczenie dla parametrów tego typu przełącznika, w szczególności izolacji.

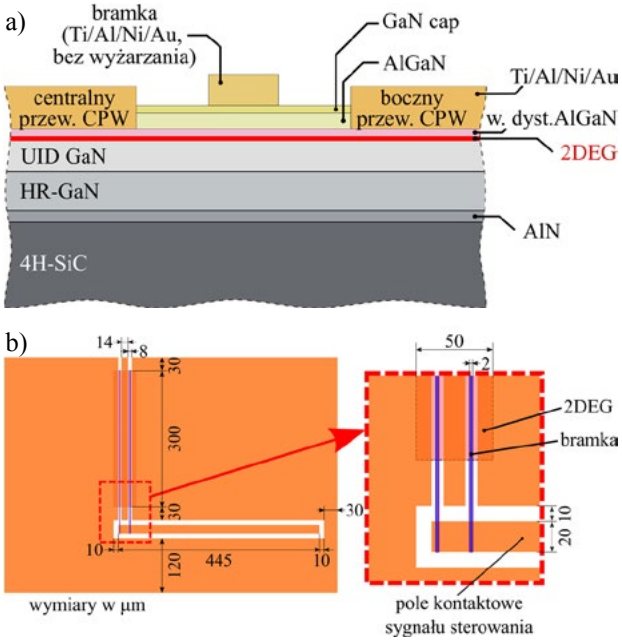


Rys. 1. Budowa przełącznika o topologii szeregowej

Wyniki badań przedstawione w [7] wskazują, że przełączniki o topologii szeregowej, choć często stosowane w zakresie mikrofalowym nie są w stanie pracując samodzielnie zapewnić dobrych parametrów w zakresie fal milimetrowych. Jest to spowodowane głównie występowaniem znacznych sprzężeń oraz niedopasowaniem impedancji związanym ze znaczącą rezystancją kanału tranzystora. Z tego względu zaprojektowano i zbadano także przełączniki o topologii bocznikowej [8].

2.2. Przełącznik o topologii bocznikowej

W konstrukcji przełączników o topologii bocznikowej wykorzystano analogiczną strukturę półprzewodnikową jak w przełączniku szeregowym. W topologii bocznikowej element aktywny umieszczony jest w taki sposób, by jego przewodzenie powodowało powstanie zwarcia w linii transmisyjnej, a co za tym idzie zablokowanie transmisji (przełącznik odbiciowy). W zaproponowanej konstrukcji struktura tranzystora została umieszczona w szczelinach linii CPW, po obu stronach przewodnika centralnego (rys. 2). Elektrody sterujące zostały umieszczone na środku szczeliny i dołączone z jednej strony do wspólnego pola kontaktowego odseparowanego od linii transmisyjnej.



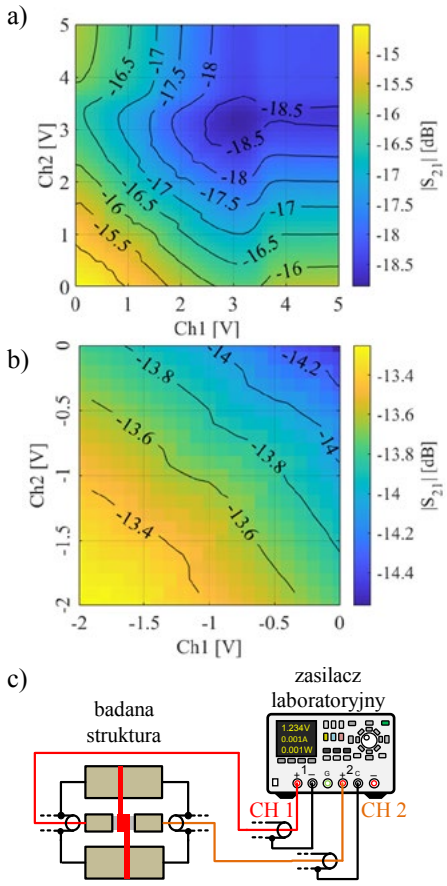
Rys. 2. Budowa przełącznika o topologii bocznikowej a) przekrój poprzeczny, b) rzut z góry

Aby umożliwić zbadanie wpływu zastosowania bramki grafenowej wykonano dwa rodzaje przedstawionego wyżej przełącznika różniące się jedynie materiałem, z którego wykonano elektrodę bramki. W wariancie z elektrodą grafenową zastosowano warstwę grafenu wytworzonego metodą CVD. W wariancie z elektrodą metalową została ona wykonana z przewodnika Ti/Al/Ni/Au nie poddawanego procesowi wyżarzania, w przeciwieństwie do kontaktów omowych.

3. BADANIA EKSPERYMENTALNE

3.1. Badania wpływu polaryzacji bramki grafenowej na parametry przełącznika

W eksperymencie wykorzystano przełącznik o topologii szeregowej opisany w punkcie 2.1. W tej konstrukcji grafenowa elektroda bramki dołączona jest stale do bocznych przewodników linii CPW, natomiast kanał tranzystora włączony jest pomiędzy dwie części przewodnika centralnego. Sygnał sterujący podawany jest tu wraz z sygnałem w.cz.. Przy braku napięcia sterującego tranzystor przewodzi umożliwiając transmisję sygnału. Wyłączenie przełącznika następuje poprzez podanie na przewód centralny CPW napięcia dodatniego, co skutkuje spolaryzowaniem złącza bramki w kierunku zaporowym. Badany przełącznik został dołączony do wektorowego analizatora obwodów za pośrednictwem sond typu GSG oraz głowic falowodowych umożliwiających pomiar na częstotliwości 100 GHz. Napięcia sterujące dołączone zostały do obu wrót przełącznika zgodnie ze schematem przedstawionym na rys. 3c. Na rys. 3a przedstawiono zmierzone charakterystyki współczynnika transmisji pomiędzy wrotami przełącznika w zależności od podanych napięć sterujących. Jak można zauważyć największą amplitudę zmian współczynnika transmisji uzyskano przy zasilaniu symetrycznym. Jest to spowodowane brakiem możliwości prawidłowego wysterowania części obszaru kanału przy zasilaniu jednostronnym w sytuacji wzrostu rezystancji kanału. Do pełnego wyłączenia przełącznika wymagane jest napięcie ok. 3,1 V. Uzyskane wyniki wskazują na stosunkowo wysokie straty i niską izolację przełącznika szeregowego. Biorąc pod uwagę wyniki badań innych wariantów przełączników szeregowych [7] można stwierdzić, że topologia ta nie jest odpowiednia dla zastosowań w zakresie fal milimetrowych.

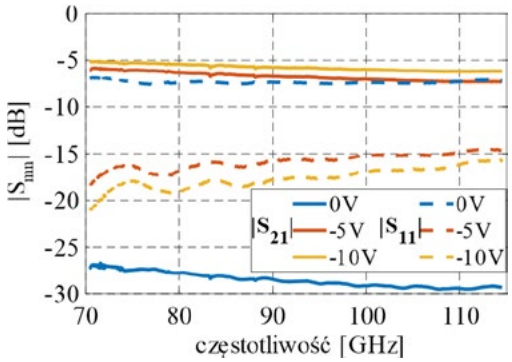


Rys. 3. Zmierzone charakterystyki współczynnika transmisji przełącznika szeregowego względem a) dodatnich, b) ujemnych napięć sterujących przy częstotliwości 100 GHz. c) Sposób dołączenia napięć sterujących

W badanym przełączniku elektrodę bramki wykonano z grafenu wytwarzanego metodą CVD, który charakteryzuje się przewodnictwem typu p. W związku z tym polaryzacja bramki w kierunku przewodzenia (ujemne napięcie sterujące) powoduje spadek koncentracji nośników (dziur) w grafenie, a co za tym idzie spadek przewodności bramki. Na rys. 3b przedstawiono zmierzone charakterystyki współczynnika transmisji dla ujemnych napięć sterujących. Jak można zaobserwować symetryczne obniżanie napięcia sterującego z obu stron przełącznika powoduje niewielki, lecz zauważalny wzrost współczynnika transmisji. Pozwala to stwierdzić, że część strat występujących w przełączniku jest spowodowane obecnością elektrody bramki. Sygnał w.cz. przedostaje się do bramki poprzez jej pojemność, a następnie jest poprzez nią zwierany do masy linii transmisyjnej. Obniżenie przewodności bramki grafenowej pozwala zatem na częściową redukcję strat w przełączniku.

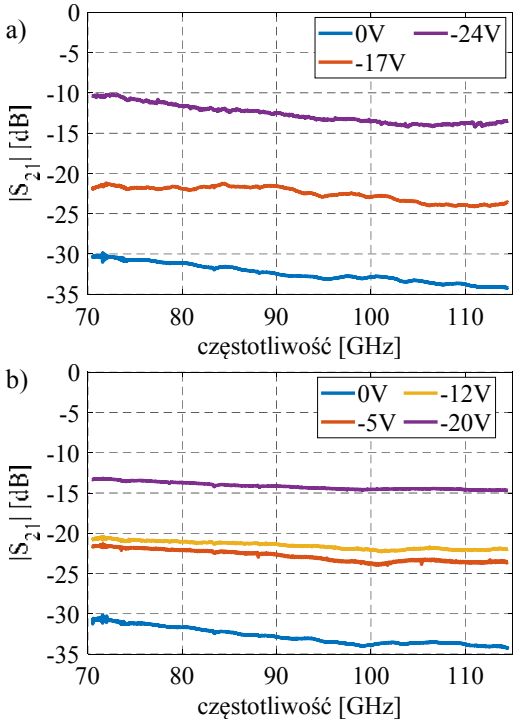
3.2. Badania przełączników o topologii bocznikowej

W badanych przełącznikach o topologii bocznikowej kanał tranzystora włączony jest pomiędzy przewodniki linii CPW. Z tego powodu do włączenia przełącznika konieczne jest wyłączenie tranzystora ujemnym napięciem podawanym na elektrodę bramki. Na rys. 4 przedstawiono zmierzone charakterystyki współczynnika transmisji przełącznika z metalową elektrodą bramki. Jak można zauważyć przełącznik wykazuje znacznie niższe straty oraz wyższą izolację w porównaniu z przełącznikiem szeregowym. Ponadto charakterystyki nie ulegają znaczącym zmianom przy górnej granicy pasma pomiarowego, co sugeruje możliwość pracy także przy wyższych częstotliwościach. Przełączniki tego typu wykazują także niskie straty odbiciowe [8]. Straty na poziomie 6 – 7 dB wynikają w dużej mierze z niskiej szerokości linii CPW, jednak jest ona konieczna dla uzyskania niskiej rezystancji kanału w stanie wyłączonym.



Rys. 4. Zmierzone charakterystyki modułów współczynników transmisji i odbicia przełącznika o topologii bocznikowej z bramką metalową

Na rys. 5 przedstawiono wyniki pomiarów charakterystyk współczynnika transmisji dwóch egzemplarzy przełącznika z bramką grafenową. W obu przypadkach osiągnięto poprawę izolacji o ok. 4 – 5 dB w stosunku do przełącznika z bramką metalową, natomiast znaczącemu wzrostowi uległ poziom strat. Należy przy tym zauważyć, że do włączenia przełącznika konieczne było zastosowanie ponad dwukrotnie wyższego napięcia sterującego niż w przypadku zastosowania bramki metalowej. W połączeniu ze znacznymi różnicami tłumienia i wymaganych napięć sterujących pomiędzy dwoma egzemplarzami wskazuje to na występowanie przypadkowych defektów produkcyjnych w warstwie grafenu. Uniemożliwią one kontrolę kanału tranzystora na całej długości bramki, uniemożliwiając całkowite włączenie przełącznika. Częściowo udaje się to skompensować wyższym napięciem sterującym powodującym zwiększenie natężenia pola elektrycznego wypychającego gaz elektronowy z kanału.



Rys. 5. Zmierzone charakterystyki modułu współczynnika transmisji przy różnych napięciach sterujących przełącznika o topologii bocznikowej z bramką grafenową: a) egzemplarz 1, b) egzemplarz 2

Długa i wąska elektroda bramki jest elementem wrażliwym nawet na niewielkie defekty. W przypadku warstw metalowych, dobrze dopracowane procesy technologiczne pozwalają na osiągnięcie odpowiednio wysokiej dokładności. Stosowane obecnie procesy wytwarzania elementów grafenowych na strukturach półprzewodnikowych wymagają trzech etapów. Warstwa grafenu wytwarzana jest na metalowej folii z której transferowana jest na docelową strukturę, a następnie trawiona. Proces transferu powoduje duże ryzyko powstawania drobnych defektów. Jednocześnie brakuje szybkich i dokładnych metod kontroli jakości warstw. Dla uzyskania dobrych parametrów układów i powtarzalności uzyskiwanych efektów konieczne jest zatem udoskonalenie procesów wytwarzania elektrod grafenowych. Korzystne może być także zoptymalizowanie konstrukcji przełącznika pod kątem większej odporności na występowanie defektów elektrody bramki np. poprzez zwiększenie jej powierzchni.

4. PODSUMOWANIE

W artykule przeanalizowano możliwości zastosowania grafenowych elektrod sterujących w przełącznikach na zakres fal milimetrowych. Zaprojektowane i zbadane struktury umożliwiły wykazanie możliwości wykorzystania właściwości grafenu do ograniczenia strat oraz poprawy izolacji przełączników. Wskazano także na korzyści z zastosowania topologii bocznikowej.

Na podstawie przeprowadzonych badań można stwierdzić, że zastosowanie elektrod grafenowych w wysokoczęstotliwościowych przyrządach półprzewodnikowych niesie za sobą istotny potencjał poprawy niektórych parametrów. Należy przy tym zauważyć, że stosowane procesy technologiczne wymagają udoskonalenia, aby możliwe było uzyskanie powtarzalnych, wysokiej jakości warstw grafenowych.

LITERATURA

[1] Bogdan Grzegorz, Godziszewski Konrad, Yashchyshyn Yevhen, Kim Cheol Ho, Hyun Seok Bong. 2020. “Time-Modulated Antenna Array for Real-Time Adaptation in Wideband Wireless Systems-Part I: Design and Characterization.” *IEEE Trans. on Antennas and Prop.* 68 (10): 6964–72.

[2] Sobolewski Jakub, Yashchyshyn Yevhen. 2022. “State of the Art Sub-Terahertz Switching Solutions.” *IEEE Access* 10: 12983–99.

[3] Campbell Charles F., Balistreri Anthony, Kao Ming Yih, Dumka Deep, Hitt John. 2012. “GaN Takes the Lead.” *IEEE Microwave Magazine* 13 (6): 44–53.

[4] Theofanopoulos Panagiotis C., Trichopoulos Georgios C.. 2019. “Modeling of Sub-Millimeter Wave Coplanar Waveguide Graphene Switches.” In *2019 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting, APSURSI 2019*, 1527–28.

[5] Zhang An Qi, Liu Zhen Guo, Wei-Bing Lu, Chen Hao. 2019. “Graphene-Based Dynamically Tunable Attenuator on a Coplanar Waveguide or a Slotline.” *IEEE Transactions on Microwave Theory and Techniques* 67 (1): 70–77.

[6] Sharma Pankaj, Perruisseau-Carrier Julien, Moldovan Clara, Ionescu Adrian Mihai. 2014. “Electromagnetic Performance of RF NEMS Graphene Capacitive Switches.” *IEEE Transactions on Nanotechnology* 13 (1): 70–79.

[7] Yashchyshyn Yevhen, Bajurko Paweł, Sobolewski Jakub, Sai Pavlo, et al. 2021. “Graphene/AlGaN/GaN RF Switch.” *Micromachines* 12 (11): 1343.

[8] Yashchyshyn Yevhen, Bajurko Pawel, Sobolewski Jakub, Sai Pavlo, Rumyantsev Sergey, and Cywinski Grzegorz. 2022. “AlGaN/GaN Schottky Barrier Single-Pole Single-Throw RF Switch.” *2022 24th International Microwave and Radar Conference, MIKON 2022*.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

StegoBackoff: Tworzenie kanału skrytego z użyciem procedury backoff w sieciach standardu IEEE 802.11

StegoBackoff: Creating a Covert Channel Using Backoff Procedure in IEEE 802.11 Networks

Geovani Teca[1]; Marek Natkaniec[2];

[1] AGH University of Krakow, Krakow, teca@agh.edu.pl
[2] AGH University of Krakow, Krakow, natkanie@agh.edu.pl

DOI: 10.15199/59.2023.4.33

Streszczenie: Kanał skrytej komunikacji to technika służąca do ukrywania tajnych wiadomości wewnątrz jawnej transmisji, stosowana do zapewnienia bezpiecznej komunikacji w niezaufanych środowiskach. W artykule przedstawiono nowy kanał skrytej komunikacji oparty na procedurze backoff sieci standardu IEEE 802.11. Artykuł opisuje implementację i dostarcza metryki do oceny wydajności pracy. Wyniki pokazują, że rozmiar ramki, wielkość ruchu oferowanego oraz liczba stacji w sieci bezprzewodowej bezpośrednio wpływają na wydajność proponowanego skrytego kanału.

Abstract: A covert channel is a technique used to conceal secret messages within an explicit transmission, applied to secure communication in untrusted environments. This paper presents a novel covert channel based on the backoff procedure in the 802.11 networks. The paper describes the implementation and provides metrics for evaluating its performance. The results reveal that the payload size, offered load and the number of stations competing for access to the wireless channel directly impact the covert channel's performance.

Słowa kluczowe: kanały skryte, mechanizm Backoff, sieci standardu IEEE 802.11, steganografia sieciowa

Keywords: covert channels, Backoff mechanism, IEEE 802.11 networks, network steganography

1. INTRODUCTION

IEEE 802.11 is a set of standards for Wireless Local Area Networks (WLANs) developed by the Institute of Electrical and Electronics Engineers (IEEE) and Commonly known as Wi-Fi [1]. Many features contributed to 802.11 networks being quickly adopted and becoming an immediate solution for connecting devices and the Internet. Devices that support the IEEE 802.11 standards are easily portable (e.g. laptops, wearable devices, and cell phones). Wi-Fi is simple to set up, either by configuring an Access Point (AP) or pressing the button on a device to start a hotspot. It can be installed in locations without existing wired network infrastructure support. Finally, expanding the network range involves incorporating additional APs, enabling seamless connectivity and uninterrupted user mobility throughout the network.

The devices in the IEEE 802.11 network share the transmission medium. A station equipped with IEEE 802.11 radio, monitoring the channel, can detect when transmission occurs by measuring the channel signal energy levels. The IEEE 802.11 implements packet encryption to prevent the data from being exposed in the channel. However, over time, vulnerabilities in IEEE 802.11 encryption protocols have been discovered [2], [3], posing security risks to networks and their users. Therefore, IEEE 802.11 network researchers must focus on developing mechanisms that allow users to send data without raising awareness that transmission is occurring.

Steganography, an ancient practice, has been adopted into the IEEE 802.11 network to enable users to send secret messages that go undetected through the shared channel. This is achieved by creating covert channels that decouple the encrypted data frame from the secret data. The explicit data frame serves as a cover or envelope for the secret message, allowing it to be transmitted without detection.

The paper introduces a novel covert channel based on the IEEE 802.11 backoff procedure and presents its implementation and associated metrics. The article makes several contributions to the field of IEEE 802.11 network steganography. Firstly, it presents a new approach to covert channel implementation using the backoff mechanism. Secondly, it highlights the significant impact of frame size, number of stations, and offered load on the covert channel's throughput, delay, jitter, and efficiency. Lastly, the paper reports the simulation results from a network based on the IEEE 802.11ax specification, which enables achieving a noteworthy throughput for the covert channel.

The organization of the research paper is as follows: Section 2 presents the related works. It shortly discusses the existing covert channels for IEEE 802.11 networks. Section 3 describes the exponential backoff mechanism used to create the covert channel. Section 4 explains the covert channel concept and operation. Section 5 presents the simulation results. Finally, Section 6 concludes the paper.

2. RELATED RESEARCHES

The IEEE 802.11 standard implementation provides numerous opportunities for creating covert channels. One approach involves utilizing fields that allow custom values. For example, the study [4] demonstrates two covert channels that can be implemented in the Sequence Control and Initial Vector (IV) field of the IEEE 802.11 MAC

header. Another study [5] presents a covert channel that inserts the secret message in the Protocol Version field of the IEEE 802.11 MAC header. Furthermore, the paper [6] describes a covert channel in IEEE 802.11e that uses the reserved bit combinations of three subfields (QoS, CF-Pollable, and CF-Poll Request) of QoS capability in the Association Request to signal the start and finish of the covert communication, and when the covert channel is established the covert message is inserted in the TXOP (Transmission Opportunity) and TID (Traffic Identifier) fields.

Another technique for creating a covert channel involves encoding the secret message in the time difference between consecutive transmissions. One example of this approach is the Covert DCF [7], which manipulates the backoff values in the Distributed Coordinated Function (DCF) to encode secret messages. The Covert-DCF implementation using an Off-The-Shelf wireless card is described in [8]. The paper [9] proposes a ternary timing covert channel. This method involves the sender and receiver monitoring the wireless channel to determine the free time intervals, which are then statistically analyzed to determine the free time interval distribution of the channel. The sender and receiver use this information to conceal and decode secret messages. The study [10] presents an IEEE 802.11 covert channel that utilizes the interarrival times of either the Beacon frames or the Probe Request frames to encode the message. The method involves changing the generation interval of the frames based on the bit combination that the sender intends to transmit.

Moreover, another technique to conceal secret data involves manipulating the physical layer during the modulation process. One such method is based on the constellation shaping modulation technique, as described in [11], [12]. In this approach, the hidden message is concealed as a distorted constellation that results from hardware and channel imperfections or noise.

To the best of our knowledge, this is the first covert channel to encode secret messages based on the backoff slot parity using the latest IEEE 802.11ax amendment. In contrast to the approach presented in [7], our method eliminates the necessity for the sender to manipulate the drawn backoff slots, guaranteeing fairness in the backoff procedure.

3. IEEE 802.11 BACKOFF PROCEDURE

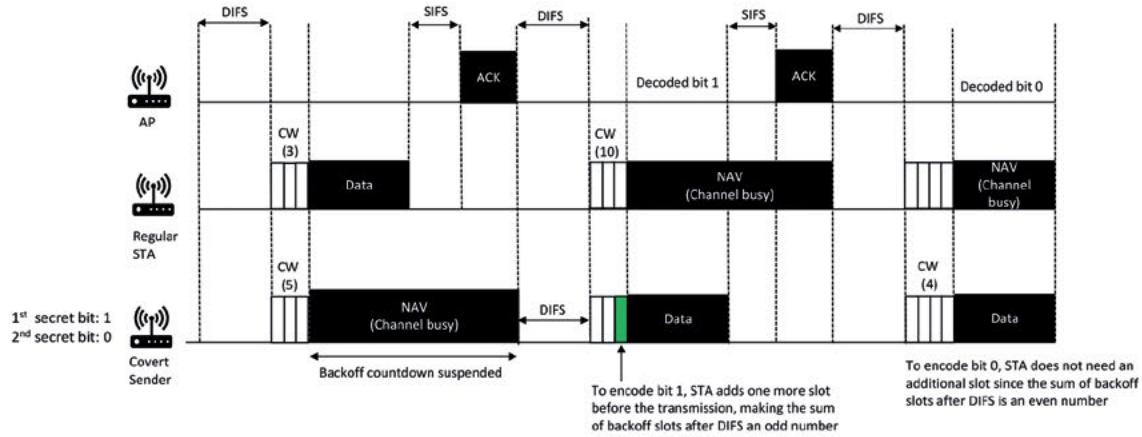
The IEEE 802.11 standards require the Distributed Coordination Function (DCF) to enable asynchronous channel access for devices sharing the transmission medium [13]. DCF employs two methods of carrier sensing: physical and virtual. A station measures the channel signal energy in physical carrier sensing to determine whether it falls below a predefined threshold. Virtual carrier sensing, on the other hand, relies on the Network Allocation Vector (NAV) of the station, which provides information about the duration of other transmissions on the channel. The DCF procedure enables asynchronous channel access for devices sharing the transmission medium. In DCF, a station that wants to send a frame must first sense the channel. If the channel is busy due to ongoing transmission by another station, the sender waits until the end of the current transmission plus a specific time known as DCF Inter Frame Space (DIFS). After DIFS, the station selects a random integer backoff slot from 0 to a value of CW (Contention Window), calculated according to formula 1, where the values of x are specific to the IEEE 802.11 amendment. Initially, every station sets CW to CW_{min}, but upon collision, it doubles until it reaches CW_{max}. However, after every successful transmission attempt, the CW is reset to CW_{min}. After drawing the backoff, the station starts the countdown until it reaches zero. At this point, the station has won the contention to access the channel. During this time, other stations set the duration of the ongoing transmission in their NAV and suspend their backoff countdown. The time interval between frames belonging to the same transmission (such as Data and ACK frames) is called Short Inter Frame Space (SIFS). After a transmission plus a DIFS period ends, the stations whose backoff has not reached zero resume their backoff countdown. The station that wins the contention for access to the channel gets a new backoff slot.

$$CW = 2^x - 1 \tag{1}$$

4. STEGOBACKOFF CONCEPT AND OPERATION

The backoff procedure allows stations to determine the number of time slots that have elapsed before a transmission can start, using the DIFS and the remaining backoff time. By exploiting this property, a covert channel can be created by encoding a secret message in the parity of the backoff slot. The bit 0 corresponds to an even backoff slot, while an odd backoff slot is used for transmitting a bit 1. This covert channel has a bandwidth of one bit per frame, and any node on the channel can be the recipient, whether it is a station or an AP. The receiver monitors the channel and counts the number of slots passed from the DIFS until a frame is transmitted from the covert station. Subtracting the DIFS from the frame arrival time gives the number of backoff slots the sender waited before transmitting the data. Based on the parity of this number, the receiver can decode the transmitted bit as a 0 or 1. The covert channel is highly resistant to steganalysis because it does not modify how backoff values are generated. If necessary, one extra slot is added to ensure the correct slot parity based on the secret message to be delivered.

To illustrate how this covert channel operates in practice, please consider Fig. 1, where a covert sender has a queue of secret messages consisting of two bits, with the first being 1 and the second being 0. After DIFS, the sender draws a random backoff of 5 slots and begins the countdown. Meanwhile, another regular station also has a frame to transmit and competes for access to the channel. It draws a lower backoff of 3 slots and wins the contention, causing the covert station to suspend its backoff countdown timer and wait until the regular station's transmission is complete. After the regular station's transmission and another DIFS interval, the covert station resumes its backoff countdown with only two slots remaining. When the backoff countdown reaches zero, the covert station waits for one more slot, making the total number of slots after DIFS odd since it intends to send bit 1. The covert stations draw a new backoff of four when the first secret data transmission finishes. After DIFS, it wins the contention as it has a lower backoff value than the regular stations and sends the frame without delay since the total number of slots passed after DIFS is even, and it intends to send a bit 0. The sender and receiver algorithms are presented in Fig.2 and Fig.3.



Rysunek 1: StegoBackoff operation

Algorithm 1: Sender procedure

```
backoff ← random(0, CW);
covert_bit ← message[i];
if covert_bit is 0 then
    if backoff is odd then
        backoff ← backoff + 1;
    end
else
    if backoff is even then
        backoff ← backoff + 1;
    end
end
```

Rysunek 2: Pseudocode for the stegobackoff sender

Algorithm 2: Receiver procedure

```
DIFS ← SIFS + 2 * Slot_time;
slots_lapsed ← (DIFS − frame_arrival_time)/Slot_time;
if slots_lapsed is even then
    decode bit 0;
else
    decode bit 1;
end
```

Rysunek 3: Pseudocode for the stegobackoff receiver

5. STEGOBACKOFF PERFORMANCE EVALUATION

5.1 Simulation Environment

The covert channel has been implemented in the ns-3 network simulator [14], a powerful discrete-event network simulator used to model and simulate various types of networks. This open-source simulator is written in C++ and Python and provides many features, including support for simulating IEEE 802.11 networks. Table 1 shows the main parameters set during the simulation. In all figures, the error of each simulation point for the 95% confidence interval did not exceed ± 2%.

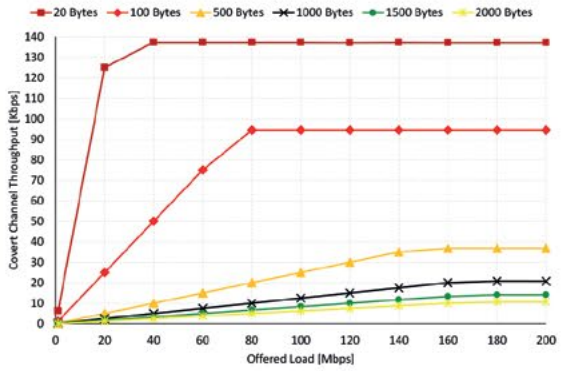
Tabela 1: Simulation Parameters

Parameter	*Value*
IEEE specification	802.11ax
Transport protocol	UDP
Channel band	2.4 [GHz]
Channel width	40 [MHz]
Guard interval	400 [μs]
Time slot	9 [μs]
SIFS	16 [μs]
DIFS	34 [μs]
Time slot	9 [μs]
MCS index	9
Mobility model	Constant
RTS/CTS	Disabled
Number of Tx and Rx antennas	1

5.2 First Scenario - Non-Competitive Channel Access

The isolated scenario involves only the covert STA and the AP, with no other STAs present. In this scenario, the covert STA generates UDP traffic, and the experiment involves increasing the offered load while maintaining a constant frame size. The experiment is repeated with varying frame sizes to assess the impact of frame size on the throughput of the covert channel as the offered load increases.

Fig. 4 shows the simulated results of the isolated scenario. The findings indicate that the payload size directly impacts the covert channel throughput. As the figure shows, smaller payload sizes generate higher covert throughput values, while larger payload sizes result in lower throughput values. This is due to the fact that a smaller payload occupies the channel for a shorter duration, providing the covert stations with more transmission opportunities within the overall transmission time. Furthermore, our observations indicate that the payload size impacts the network saturation threshold. In particular, the network reaches its saturation point for small payload sizes with less offered load, while larger payload sizes require more load to achieve saturation. This can be attributed to the fact that as the payloads become heavier, more traffic must be generated to reach or exceed the threshold.

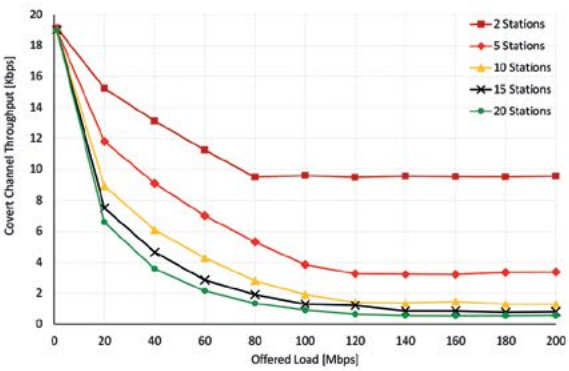


Rysunek 4: Covert channel throughput vs. offered load for different payload sizes, with 1 STA and 1 AP

5.3 Second Scenario - Contention-Based Channel Access

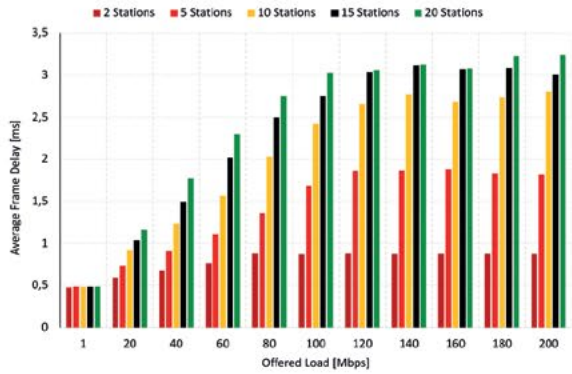
In second scenario, we introduce additional regular stations to the network, gradually increasing the network's density. The covert STA maintains a constant frame size of 1024 bytes and offers a load of 200 Mbps. The experiment aims to investigate the effect of increasing traffic generated by additional stations on the performance of the covert channel, including throughput, delay, jitter, and frame efficiency, which is the ratio of frames received over frames sent in percentage.

The simulation results in Fig. 5 show the impact of adding 1, 5, 10, 15, and 20 more stations to the network. It is observed that when the external traffic is low, at 1 Mbps, the covert channel is not affected and achieves approximately 19 Kbps throughput for all five experiments. This can be attributed to the fact that the covert station generates much higher traffic (200 Mbps) than 1 Mbps and, thus, do not face significant interference. However, as the offered load increases, the impact of external traffic becomes noticeable, and the covert throughput starts to decay. This is due to the increased contention for channel access, as more data is to be sent by external stations. Furthermore, it is observed that after adding 15 stations, the throughput values are close, suggesting that adding more stations beyond this point will have a minimal impact on the already decayed network throughput.



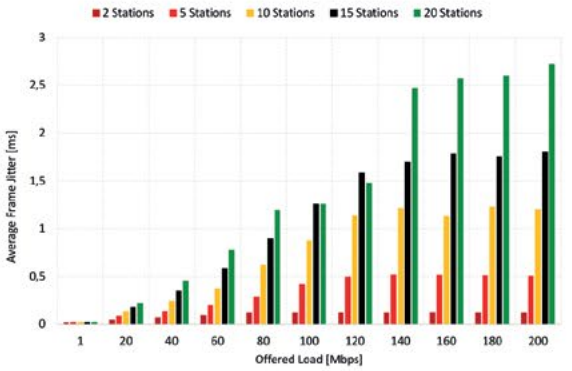
Rysunek 5: Covert channel throughput vs. regular stations offered load

Fig. 6 illustrates the impact of external stations on the delay of the covert channel. The results show that the delay increases proportionally to the offered load from the regular stations. As the offered load increases, the delay growths due to congestion. When traffic is high, more collisions occur, and the receiving queue at the AP increases, causing processing delays. Therefore, the delay experienced by the covert channel is directly proportional to the amount of offered load. Overall, in the worst-case scenario, with 20 stations generating 200 Mbps of traffic, the covert channel delay is below 3,5 ms, regarded as an excellent outcome.



Rysunek 6: Covert channel delay vs. regular stations offered load

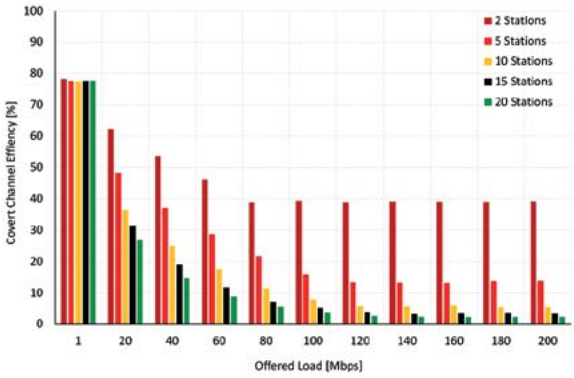
As shown in Fig. 7, a correlation between the offered load and frame jitter is similar to the relationship with delay. Initially, even with varying numbers of stations in the network, a minimal jitter (around 0.1 ms) is observed, indicating that low traffic has an insignificant impact on the covert channel. Even with a significant traffic load of 60 Mbps, the jitter remains below 1 ms, which is considered acceptable. Only in the scenario with 20 stations and from 160 Mbps of traffic does the jitter exceed the threshold of 2 ms. These findings suggest that the covert channel maintains acceptable jitter levels under diverse scenarios.



Rysunek 7: Covert channel jitter vs. regular stations offered load

The efficiency of the covert channel is highly dependent on the offered load, as demonstrated by the results presented in Fig. 8. This figure displays the percentage of frames successfully delivered to AP and the percentage lost depending on the number of stations. The efficiency of the covert channel is inversely proportional to the offered load, as observed. At a low offered load of 1 Mbps, over 80% of frames are successfully delivered. However, as the

amount of traffic increases, there is a significant reduction in efficiency. For instance, in scenarios with 10, 15, and 20 stations, the efficiency drops below 40% when the offered load reaches 20 Mbps. As the offered load increases, the increase in frame collisions becomes more evident. This decrease in efficiency is mainly due to the increase in frame collisions, which is intensified by the number of stations.



Rysunek 8: Covert channel efficiency vs. regular stations offered load

6. CONCLUSION

This paper introduces a novel covert channel based on a backoff mechanism, analyzed using the ns-3 network simulator with the latest IEEE 802.11ax amendment. Our analysis investigated the impact of frame size, offered load, and the number of stations on the covert channel's throughput, delay, jitter, and frame efficiency.

Our findings revealed that achieving higher throughput requires using smaller frame sizes, as larger payloads tend to decrease throughput. Furthermore, an increase in the offered load damages the covert channel performance, and the number of stations intensifies this effect, accelerating network degradation.

We highlight that our experiments were conducted with considerably higher offered loads than those usually encountered in real-life scenarios. Despite these demanding conditions, the covert channel exhibited satisfactory performance. Based on our findings, we confidently conclude that, when employed with an appropriate payload size in a moderately congested network, the covert channel represents an excellent option for establishing a covert communication channel. It provides adequate covert throughput while maintaining acceptable levels of delay and jitter, seamlessly incorporating into the backoff procedure without causing any disruptions to normal network operations.

LITERATURA

[1] IEEE Standard for Information Technology. 2021. "Telecommunications and Information Exchange between Systems - Local and Metropolitan Area Networks–Specific Requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. IEEE Std 802.11-2020". IEEE Std 802.11-2020 (Revision of IEEE Std 802.11-2016): 1-4379.

[2] Vishal Kumkar, Akhil Tiwari, Pawan Tiwari, Ashish Gupta, Seema Shrawne. 2012. "Vulnerabilities of Wireless Security protocols (WEP and WPA2)". International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 1 (2) : 34-38.

[3] Kemal Bicakci, Bulent Tavli. "Denial-of-Service attacks and countermeasures in IEEE 802.11 wireless networks". Computer Standards & Interfaces 31 (5) : 931-941.

[4] Lilia Frikha, Zouheir Trabelsi, Wassim El-Hajj. 2008. "Implementation of a Covert Channel in the 802.11 Header". International Wireless Communications and Mobile Computing Conference, Crete, Greece : 594-599.

[5] Goncalves Ricardo. Tummala Murali. Mceachen John. 2012. "Analysis of a MAC Layer Covert Channel in 802.11 Networks". International Journal on Advances in Telecommunications 5(3&4) : 131-140

[6] Hong Zhao. 2014. "Covert channels in 802.11e wireless networks". Wireless Telecommunications Symposium, Washington, DC, USA : 1-5.

[7] Russell Holloway, Raheem Beyah. 2011. "Covert DCF: A DCF-Based Covert Timing Channel in 802.11 Networks". IEEE Eighth International Conference on Mobile Ad-Hoc and Sensor Systems, Valencia, Spain : 570-579.

[8] Sakthi V. Radhakrishnan, A. Selcuk Uluagac, Raheem Beyah. 2013. "Realizing an 802.11-based covert timing channel using off-the-shelf wireless cards". IEEE Global Communications Conference (GLOBECOM), Atlanta, GA, USA : 722-728.

[9] F. Tahmasbi, N. Moghim and M. Mahdavi, "Ternary timing covert channel in wireless 802.11". 2015. 12th International Iranian Society of Cryptology Conference on Information Security and Cryptology (ISCISC), Rasht, Iran : 73-78.

[10] T. Owens Walker and Kevin. D. Fairbanks, "An off-the-shelf, low detectability, low data rate, timing-based covert channel for IEEE 802.11 wireless networks". 2017. 14th IEEE Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA: 835-840.

[11] Grzesiak Krystian, Zbigniew Piotrowski, and Jan M. Kelner. 2021. "A Wireless Covert Channel Based on Dirty Constellation with Phase Drift". Electronics 10(6) : 647.

[12] Pengcheng Cao, Weiwei Liu, Guangjie Liu, Xiaopeng Ji, Jiangtao Zhai, Yuewei Dai, "A Wireless Covert Channel Based on Constellation Shaping Modulation". 2018. Security and Communication Networks, 2018 : 1-15.

[13] Natkaniec Marek, and Andrzej R. Pach. "An analysis of the backoff mechanism used in IEEE 802.11 networks". 2000. Proceedings ISCC 2000. Fifth IEEE Symposium on Computers and Communications, Antibes-Juan Les Pins, France : 444-449.

[14] https://www.nsnam.org

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

NEGATYWNY WPŁYW NISKICH POZIOMÓW MOCY NADAWANIA NA WYDAJNOŚĆ PRACY SIECI BEZPRZEWODOWYCH STANDARDU IEEE 802.11ax.

NEGATIVE IMPACT OF LOW TRANSMISSION POWER LEVELS ON THE PERFORMANCE OF IEEE 802.11ax WIRELESS NETWORKS.

Szymon Stryczek[1]; Marek Natkaniec[2]

[1] Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Kraków, szst@agh.edu.pl
[2] Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Kraków, natkanie@agh.edu.pl

. . .

DOI: 10.15199/59.2023.4.34

Streszczenie: W artykule opisano problemy związane z wydajnością pracy sieci standardu IEEE 802.11ax z włączonym mechanizmem kolorowania obszarów BSS przy użyciu niskich mocy nadawania. Autorzy przedstawiają scenariusze symulacyjne, efektem których są obserwacje niepokojącego działania sieci bezprzewodowych przedstawione na wykresach. Wskazano inne prace z zakresu kolorowania sieci, także te dowodzące problemów występujących w sieciach WLAN. W artykule odniesiono się do kwestii dobierania wartości parametrów standardu IEEE 802.11ax.

Abstract: The article describes issues related to the performance of IEEE 802.11ax network operation when the BSS coloring mechanism is enabled with low transmit power. The authors present simulation scenarios that result in observations of concerning behavior of wireless networks, depicted in plots. They refer to other works on network coloring, including those highlighting problems occurring in WLANs. The article addresses the issue of parameter selection in IEEE 802.11ax standard.

Słowa kluczowe: IEEE 802.11ax, moc nadawania, próg OBSS_PD, spadek wydajności pracy

Keywords: IEEE 802.11ax, transmission power, OBSS_PD threshold, throughput degradation

1. WSTĘP

W niniejszej pracy podjęto temat analizy problemu spadku wydajności działania sieci bezprzewodowych wspierających kolorowanie obszarów BSS. Przeprowadzone badania wskazują, że analizowane spadki wydajności są ściśle powiązane z ustawioną wartością progu OBSS_PD (Overlaping Basic Service Set Packet Detection) oraz mocą nadajnika. Standard sieci bezprzewodowych IEEE 802.11ax [1] wprowadza kolorowanie obszarów sieci bezprzewodowych poprzez mechanizm OBSS (Overlaping Basic Service Set). Udostępnione zostały 64 możliwości ustawienia koloru obszaru, w zakresie od 0 do 63. Kolor oznaczony jako 0 uważany jest za obszar bez nadanego koloru. Zatem oferowane są tylko 63 odrębne identyfikatory obszarów sieci Wi-Fi. Zdefiniowano dodatkowy wektor alokacji zasobów dla ruchu w ramach obszaru BSS (Basic Service Set) oznaczonego dedykowanym kolorem, nazwanym Intra-BSS NAV [1]. Kanał uznawany jest za wolny, gdy oba wektory NAV przyjmują wartość zero. Standardowy wektor NAV nazwany Basic NAV jest uaktualniany po rozpoznaniu ramki sieci Wi-Fi o poziomie mocy przekraczającej próg OBSS_PD, z tak zwanym międzyobszarowym PPDU (Physical Protocol Data Unit) albo PPDU bez nadanego identyfikatora obszaru. Intra-BSS NAV jest uaktualniany zawsze, gdy stacja znajduje się w zasięgu innej stacji pracującej w określonym obszarze z nadanym tym samym kolorem, a poziom mocy sygnału przekracza próg wykrywania nośnej. Motywacją prowadzonych badań jest potrzeba zweryfikowania wydajności działania sieci standardu IEEE 802.11ax w scenariuszu opartym o zmianę poziomu mocy nadajnika. Celem tych badań jest opracowanie technik zarządzania mocą nadajnika, aby zminimalizować interferencje sygnałów radiowych sieci pracujących na ograniczonym obszarze.

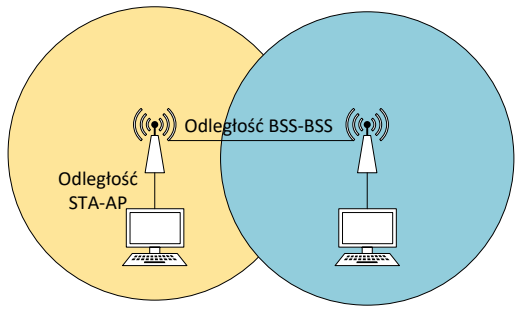
2. PRZEGLĄD LITERATURY

Problem kolorowania sieci i dobierania progów OBSS_PD jest zagadnieniem szeroko opisanym w literaturze. Autorzy [2] wskazali zysk z kolorowania sieci w przypadku dwóch obszarów. Obserwowana poprawa przepływności z wykorzystaniem kolorowania osiąga nawet 20% w stosunku do scenariusza, kiedy kolorowanie jest nieaktywne. Wykazano większe różnice w przypadku rosnącej liczby stacji podłączonych do jednego punktu dostępowego. Analiza [3] przedstawia topologię trzech obszarów ustawionych w linii z wieloma podłączonymi stacjami. Wyniki symulacyjne w oparciu o symulator ns-3 wskazują zysk około 5% w zewnętrznych obszarach i 47% w środkowym obszarze dla sieci z włączonym kolorowaniem obszarów względem scenariusza z nieaktywnym kolorowaniem. Jest to znacząca poprawa powiązana z wykorzystaniem techniki kolorowania obszarów. Innym przykładem wskazania zysku dla sieci gęstych z użyciem techniki kolorowania BSS są wyniki uzyskane w [4]. W kolejnym artykule również przedstawiono, że w przypadku rosnącej liczby urządzeń w sieci kolorowanie obszarów znacząco wpływa na poprawę wydajności sieci. W analizowanej topologii na obszarze 1,85 km2 rozmieszczono 127 punktów dostępowych i 1431 stacji, [5]. Zmianie podlegał procentowy udział stacji pracują-

cych ze wsparciem kolorowania sieci. Wykazano, że liczba stacji wspierających kolorowanie obszarów ma pozytywny wpływ na wydajność całej sieci. W [6] przedstawiono jak istotny jest właściwy dobór parametrów sieci oraz wykazano wzrost wydajności pracy sieci w przypadku wartości progu OBSS_PD na poziomie -72,0 dBm. Zaproponowano również nowe metody dopasowania progu OBSS_PD do warunków sieciowych [7]. Negatywny wpływ kolorowania obszarów w przypadku zbyt dużych wartości progu OBSS_PD zaprezentowano w [8], gdzie krytycznym punktem okazała się wartość -64,0 dBm. W [9] zaproponowano algorytm dostosowywania progu OBSS_PD, dopasowując jego wartość do warunków radiowych. Badania skuteczności algorytmu wykonano z wykorzystaniem oprogramowania MATLAB. Efektem badań jest poprawa dostępu do kanału w sieci. Zwiększanie liczby wektorów NAV może prowadzić do pogorszenia parametru opóźnień oraz przepływności w sieci, co wykazano w [10]. W oparciu o model analityczny oraz badania symulacyjne, autorzy wskazali, że dodatkowe wektory NAV mogą mieć przeciwny do założonego efekt działania. Ważnym zastosowaniem kolorowania obszarów jest możliwość zwiększenia pojemności sieci ultra-gęstych, szczególnie w obszarze współpracy między technologiami. W [11] autorzy podjęli tematykę dostosowania progów dla ruchu wewnątrz obszarowego, między-obszarowego i między dwoma różnymi technologiami, czyli Wi-Fi i LTE-U. Przeprowadzone symulacje oparte o trzy różne konfiguracje sieciowe, wskazują na duży wpływ progów ED (Energy Detect) ustawionych w sieciach Wi-Fi na średnie przepływności nakładających się obszarów, korzystających z technologii komórkowych [12]. Zmiana progów ED w sieciach LTE nie powodowała drastycznych zmian przepływności, w obszarach Wi-Fi, jak i komórkowych. Natomiast zmiana progów ED w sieciach Wi-Fi miała zauważalny wpływ na wydajność wszystkich obszarów. Wyższa wartość progów implikowała wolniejszą trasmisję w obu typach sieci.

3. PROBLEM DEGRADACJI WYDAJNOŚCI PRACY SIECI WI-FI

W celu przeprowadzenia badań rozszerzono symulator ns-3 w wersji 3.37 [13] o funkcjonalność dodatkowego wektora Intra-BSS NAV. Do tej pory symulator ns-3.37 oferował kolorowanie obszarów oraz aktualizację wartości tylko jednego wektora Basic NAV. Różnice w działaniu wersji 3.37 bez wsparcia dodatkowego wektora NAV, są bliskie rezultatom działania wersji rozszerzonej o Intra-BSS NAV. Wynika to z zastosowanej metody inkrementacji wektora Basic-NAV w ns-3.37. Wartość wektora jest zwiększana odpowiednio do zidentyfikowanego ruchu. Rozwiązanie to pozwala przypisać do jednego wektora NAV wartość odpowiadającą większej z dwóch wektorów w rozszerzonej implementacji. Takie rozwiązanie sprawia, że nietrywialnym jest odróżnienie, na podstawie porównania wyników wydajności sieci, symulatora podstawowego ns-3.37 od opracowanej przez autorów wersji rozszerzonej. Jednakże, standard [1] nie przewiduje możliwości zastąpienia dwóch wektorów NAV algorytmem inkrementacji jednego z nich. Zatem wdrożone rozwiązanie wprowadza zgodność symulatora ns-3.37 z wymaganiami standardu IEEE 802.11ax. Do obserwacji działania samej techniki kolorowania, bez inkrementacji wektora NAV, konieczne jest obniżenie progu OBSS_PD do poziomu progu wykrywania nośnej CCA ED (Clear Channel Assessment Energy Detection). Zaproponowano topologię sieciową [Rysunek 1] oraz ustalono parametry symulacji [Tabela 1]. Urządzenia sieciowe umieszczono w dwuwymiarowej przestrzeni, zmieniając w kolejnych próbach odległość między punktami dostępowymi obszarów. Stacje umieszczono w równej odległości od punktów dostępowych dla każdego obszaru. Każdy z BSS składał się z jednego punktu dostępowego i jednej podłączonej stacji. Zastosowano model kanału 'MultiModelSpectrumChannel', model strat 'FriisPropagationLossModel' oraz następujący model opóźnień: 'ConstantSpeedPropagationDelayModel'. Symulację powtarzano zmieniając odległość między stacją, a punktem dostępowym, odległość między punktami dostępowymi w obu obszarach, moce nadawania urządzeń oraz ustawienia kolorowania obszarów, w tym progu OBSS_PD.



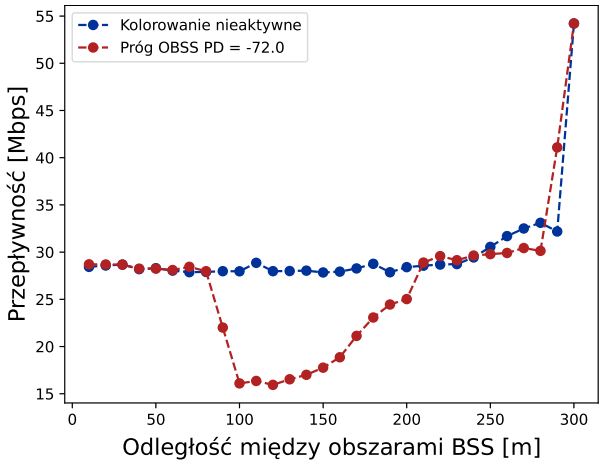
Rysunek 1: Analizowana topologia sieciowa.

Tabela 1: Parametry symulacji

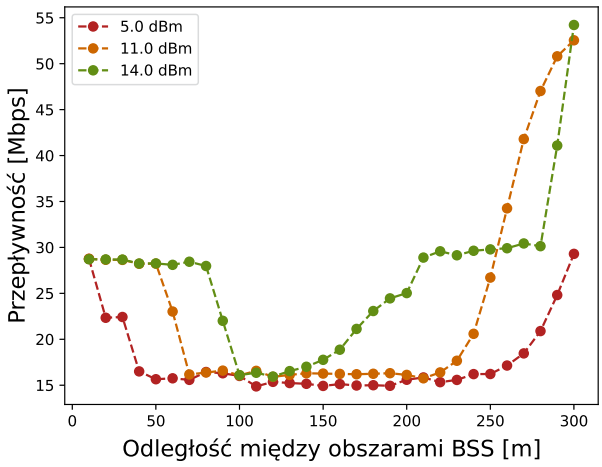
Parametr	*Wartość*
Ruch oferowany	70 Mbps
Pasmo częstotliwości	5 GHz
Szerokość kanału	20 MHz
Moc nadawania STA	5, 11, 14 dBm
Moc nadawania AP	5, 11, 14 dBm
Czułość odbiornika	-92 dBm
MCS	5
Próg CCA ED	-82 dBm
Próg CCA SD	-62 dBm
Próg OBSS_PD	od -82 do -62 dBm

Przedstawianym parametrem jest zagregowana przepływność dla obu obszarów, a ruch oferowany między stacjami i punktami dostępowymi jest generowany równo dla kierunku 'uplink' oraz 'downlink'. Założona wartość ruchu oferowanego pozwala na pracę sieci w pełnym nasyceniu w założonych warunkach. Podejście to pozwala weryfikować wpływ badanych konfiguracji na wydajność pracy wszystkich urządzeń w sieci. Zaobserwowano spadki wydajności pracy sieci z aktywnym kolorowaniem obszarów oraz ustawionym progiem OBSS_PD na wartość -72,0 dBm, względem nieaktywnej techniki kolorowania obszarów [Rysunek 2]. Zmiana parametrów odległości między stacją, a punktem dostępowym w ramach obszaru oraz regulowanie mocy nadajników powoduje, że obserwowane obniżenie wydajności pracy następuje przy innej odległości między obszarami oraz zmienia swoją charakterystykę. Na podstawie symulacji można przyjąć, że niezależnie czy mecha-

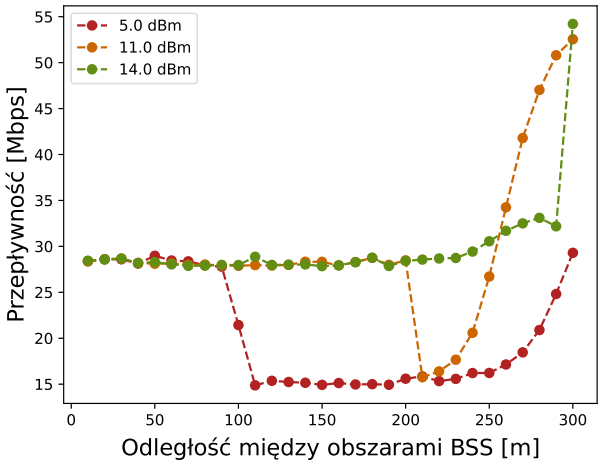
nizm OBSS_PD jest aktywny czy nieaktywny, problem ten jest ściśle powiązany z poziomem mocy nadajników. Czym mniejsza odległość między punktem dostępowym, a stacją lub czym wyższa moc nadawania tym mniejszy jest spadek wydajności [Rysunek 3, Rysunek 4]. Zwiększając odległość między punktem dostępowym lub zmniejszając moc nadawania obserwujemy pogorszenie wielkości ruchu realizowanego. Niniejsza obserwacja stała się podstawą do niżej opisanych kroków mających na celu zidentyfikowanie przyczyn obserwowanych spadków wydajności, gdyż ma to krytyczne znaczenie w kontekście zastosowań mających na celu regulowanie poziomu mocy nadajników w celu eliminacji lub minimalizacji interferencji na obszarach pracy wielu urządzeń podłączonych do punktów dostępowych sieci bezprzewodowych, zwanych sieciami gęstymi i bardzo gęstymi. Autorzy pracy uważają, że podejście badania wydajności działania sieci przy niższych poziomach nadawania i zmianach progu OBSS_PD może znacząco przyczynić się do poprawy działania gęstych sieci bezprzewodowych.



Rysunek 2: Wykres przepływności w funkcji odległości między obszarami BSS dla włączonego i wyłączonego mechanizmu kolorowania.



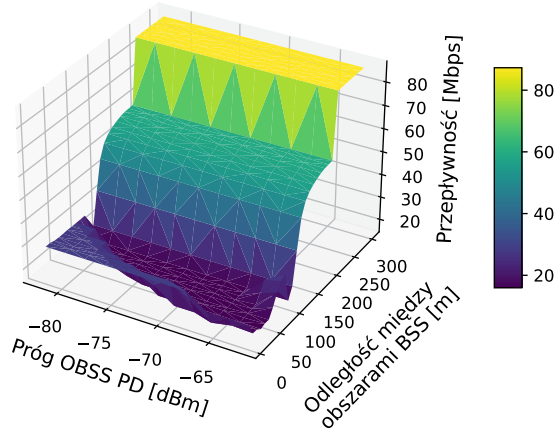
Rysunek 3: Wykres przepływności w funkcji odległości między obszarami BSS w zależności od mocy nadajnika - kolorowanie włączone, próg OBSS PD wynosi -72,0 dBm.



Rysunek 4: Wykres przepływności w funkcji odległości między obszarami BSS dla w zależności od mocy nadajnika - kolorowanie wyłączone.

4. WYNIKI BADAŃ

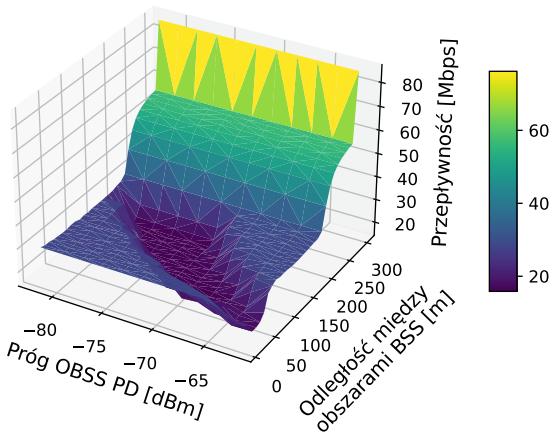
Zaobserwowany problem spadku wydajności pracy sieci Wi-Fi dla małych mocy nadawania oraz użycia kolorowania sieci pojawia się w wielu scenariuszach symulacyjnych. Symulacje prowadzono dla wielu wartości zmieniając odległość między obszarami w zakresie od 10 do 1000 metrów, odległość między stają, a punktem dostępowym w zakresie od 5 do 30 metrów, moc nadawania urządzeń od 2 do 20 dBm, oraz próg OBSS_PD w zakresie od -82 do -62 dBm. Sprawdzono także, jaki wpływ na wyniki ma aktywacja i dezaktywacja techniki kolorowania obszarów. Problem objawia się w wielu konfiguracjach symulacji. Prezentowane wyniki są efektem szerokich analiz oraz doboru takich scenariuszy, które pozwolą w sposób krytyczny przeanalizować zaistniały problem. Spadek wydajności staje się niezwykle istotny dla niskich mocy i niewielkich odległości między stacją, a punktem dostępowym [Rysunek 5].



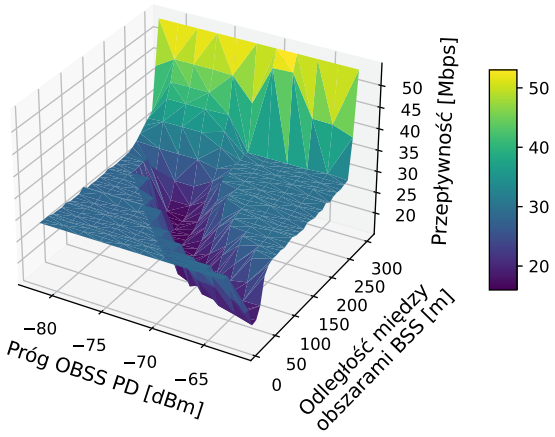
Rysunek 5: Wykres przepływności w funkcji odległości między obszarami BSS oraz progu OBSS_PD dla mocy nadajniki 5dBm i odległości STA-AP 15m.

Z przeprowadzonych badań wynika, że obniżenie wydajności pracy sieci występuje dla pełnego zakresu regulacji progu OBSS_PD. Zatem we wskazanym scenariuszu jednoznacznie stwierdzono negatywny wpływ koloro-

wania obszarów BSS oraz wsparcia dwóch wektorów NAV na wydajność pracy sieci. Prezentowane wyniki pozwalają stwierdzić, że szybkość transmisji danych w sieci drastycznie zmniejsza się, aż do całkowitego zablokowania kanału radiowego. W przypadku, który nie zakłada zmiany położenia urządzeń w sieci, dopasowanie progu OBBS_PD staje się niewystarczające. Koniecznym jest regulowanie wartości innych parametrów. Dla zwiększonego poziomu mocy zapewniana jest możliwość pracy stacji w większej odległości, jednak uzyskany wynik prowadzi do zbliżonych z powyższymi wniosków [Rysunek 6]. Efektywność działania sieci poprawia się, gdy wartość progu OBSS_PD, zbliża się, a następnie osiąga wartość progu CCA ED. Zatem, jest to sytuacja, w której sieć pracuje tak, jakby nie było wsparcia dla drugiego wektora NAV. Porównując wyniki dla odległości między stacją, a punktem dostępowym ustawioną na 25 i 30 metrów oraz odpowiednio mocą nadajników 11 dBm i 14 dBm, zaobserwowano, że w przypadku niższych poziomów mocy, jest zdecydowanie więcej możliwych ustawień odległości między obszarami, w których obserwowano spadki przepływności [Rysunek 6, Rysunek 7].

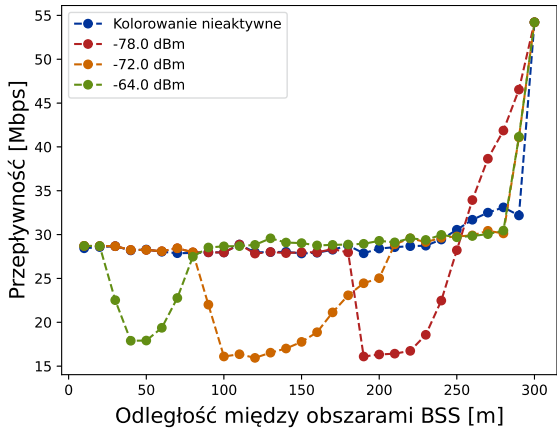


Rysunek 6: Wykres przepływności w funkcji odległości między obszarami BSS oraz progu OBSS_PD dla mocy nadajnika 11dBm i odległości STA-AP 25m.



Rysunek 7: Wykres przepływności w funkcji odległości między obszarami BSS oraz progu OBSS_PD dla mocy nadajnika 14dBm i odległości STA-AP 30m.

W sieci występują również zdarzenia, które przynoszą nieoczekiwane spadki wydajności dla bardziej złożonych konfiguracji pracy sieci. Dla odległości stacji od punktu dostępowego ustawionej na 30 m oraz poziomu mocy nadajników ustawionych na 14 dBm, nie wykazano redukcji wydajności dla nieaktywnego kolorowania obszarów oraz różne charakterystyki degradacji przepływności w zależności od progu OBSS_PD [Rysunek 8]. Zatem dokonując zmiany więcej niż jednego parametru, włącznie z całkowitą dezaktywacją mechanizmu kolorowania obszarów BSS, możliwe staje się dobranie parametrów pozwalających na pracę sieci bezprzewodowej z pełną wydajnością. W przeciwnym wypadku sieć może pracować wolniej o 50%, a w skrajnych przypadkach blokować nadawanie, ograniczając przepływność nawet poniżej 1 Mbps.



Rysunek 8: Wykres przepływności w funkcji odległości między obszarami BSS dla różnych wartości progu OBSS_PD - moc nadajnika 14dBm, odległość STA-AP 30m.

5. WNIOSKI

Zastosowanie kolorowania obszarów BSS w sieciach ze sterowaniem mocą nadajników może okazać się szkodliwe dla uzyskania wysokiej wydajności pracy sieci gęstych. Ustawienie stałych wartości progów OBSS_PD i CCA ED, bez powiązania z ustawieniami mocy nadajnika, może doprowadzić do silnych interferencji między nieprawidłowo rozpoznanymi ramkami. Urządzenia sąsiednich obszarów nie są wstanie rozpoznać komunikacji zachodzącej w obcym obszarze. Jednakże, rozpoznają zniekształcony sygnał radiowy, który może być interpretowany jako zajęty kanał na potrzeby innej komunikacji. W konsekwencji skutkuje to niewłaściwym ustawieniem wektorów alokacji NAV i wzajemnym blokowaniem się urządzeń pozostających w zasięgu konkurencyjnych transmisji radiowych. Obserwacje z prowadzonych badań prowadzą do wniosku, że w przypadku sieci gęstych sterowanie mocą nadajników dla ograniczenia interferencji między obszarami bez powiązania z innymi parametrami warstwy fizycznej i MAC jest podejściem o bardzo granicznej skuteczności. Co więcej, standard IEEE 802.11ax definiuje dopuszczalne wartości progu OBSS_PD w zależności od poziomu mocy nadawczej. Wzór (1) przedstawia górne organicznie dla progu OBSS_PD w zależności od poziomu mocy nadajnika i specyfiki użytego kanału radiowego.

$$\begin{aligned} OBSS_PD_{level} \leq max(&OBSS_PD_{min}, \\ &min(OBSS_PD_{max}, OBSS_PD_{min} \\ &+(TX_PWR_{ref} - TX_PWR))) \\ &+log_{10}(PPDU_BW/20MHz) \end{aligned} \quad (1)$$

OBSS_PDlevel to wartość progu OBSS_PD. OBSS_PDmax, OBSS_PDmin - maksymalna i minimalna wartość progu. TX_PWR oraz TX_PWRref to moc nadajnika i moc referencyjna zdefiniowana w standardzie [1]. PPDU_BW odpowiada szerokość użytego kanału radiowego.

W uzyskanych wynikach, część symulacji przekracza obliczony według definicji limit. Progi zostały ustawione w ten sposób, aby wyniki przedstawione na wykresach reprezentowały pełen zakres możliwości regulacji wartości progu OBSS_PD od OBSS_PDmin do OBSS_PDmax.

6. PODSUMOWANIE

Niniejsza praca przedstawia problem degradacji wydajności działania sieci standardu IEEE 802.11ax. Przedstawiono wpływ zmniejszania poziomu mocy nadawczej na wydajność pracy sieci WLAN z włączoną i wyłączoną techniką kolorowania BSS. Wykazano, że niższe moce nadawcze urządzeń przyczyniają się do powstawania zjawisk niepożądanych, a zatem nie mogą być stosowane bez korelacji z innymi parametrami zdefiniowanymi w standardzie. Zmiany pojedynczych parametrów pracy sieci bezprzewodowych związanych z mocą sygnału odbieranego, takich jak, moc nadajnika, odległość między urządzeniami, czy progi interpretacji sygnałów radiowych mogą mieć negatywny wpływ na działanie sieci. Zatem podejście to nie powinno być uznawane za metodę na zwiększenie wydajności pracy sieci gęstych.

LITERATURA

[1] "IEEE Standard for Information Technology–Telecommunications and Information Exchange between Systems Local and Metropolitan Area Networks–Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 1: Enhancements for High-Efficiency WLAN," in IEEE Std 802.11ax-2021, vol., no., pp.1-767, 19 May 2021, doi: 10.1109/IEEESTD.2021.9442429.

[2] A. Behara and T. G. Venkatesh, "Performance Study of High-Efficiency IEEE 802.11ax WLAN Standard Using NS-3 Simulator," 2022 IEEE 27th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), Paris, France, 2022, pp. 226-231, doi: 10.1109/CAMAD55695.2022.9966905.

[3] N. Šepić, E. Kočan, Z. Veljović and M. Pejanović, "Assessment of novel solutions for throughput enhancement in IEEE 802.11ax networks," 2019 27th Telecommunications Forum (TELFOR), Belgrade, Serbia, 2019, pp. 1-4, doi: 10.1109/TELFOR48224.2019.8971156.

[4] F. Wilhelmi, S. Barrachina-Muñoz and B. Bellalta, "On the Performance of the Spatial Reuse Operation in IEEE 802.11ax WLANs," 2019 IEEE Conference on Standards for Communications and Networking (CSCN), Granada, Spain, 2019, pp. 1-6, doi: 10.1109/CSCN.2019.8931315.

[5] I. Selinis, M. Filo, S. Vahid, J. Rodriguez and R. Tafazolli, "Evaluation of the DSC algorithm and the BSS color scheme in dense cellular-like IEEE 802.11ax deployments," 2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Valencia, Spain, 2016, pp. 1-7, doi: 10.1109/PIMRC.2016.7794832.

[6] N. Šepić, E. Kočan and M. Pejanović-Djurišić, "Evaluating spatial reuse in 802.11ax networks with interference threshold adjustment," 2020 24th International Conference on Information Technology (IT), Zabljak, Montenegro, 2020, pp. 1-4.

[7] I. Selinis, K. Katsaros, S. Vahid and R. Tafazolli, "Control OBSS/PD Sensitivity Threshold for IEEE 802.11ax BSS Color," 2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Bologna, Italy, 2018, pp. 1-7, doi: 10.1109/PIMRC.2018.8580778.

[8] M. Natkaniec, N. Bieryt, "An analysis of BSS coloring mechanism in IEEE 802.11ax dense networks," 2022 International Journal of Electronics and Telecommunications, vol. 68, No 4, pp. 855-862, Polish Academy of Sciences Committee of Electronics and Telecommunications, doi: 10.24425/ijet.2022.143894

[9] Y. Kim, G. Kim, T. Kim and W. Choi, "Transmission Opportunity-Based Distributed OBSS/PD Determination Method in IEEE 802.11ax Networks," 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, 2020, pp. 469-471, doi: 10.1109/ICAIIC48513.2020.9065000.

[10] Kwon, D.; Kim, S.-W.; Kim, J.; Mohaisen, A. Interference-Aware Adaptive Beam Alignment for Hyper-Dense IEEE 802.11ax Internet-of-Things Networks. Sensors 2018, 18, 3364. https://doi.org/10.3390/s18103364

[11] M. Hirzallah, W. Afifi and M. Krunz, "Full-Duplex Spectrum Sensing and Fairness Mechanisms for Wi-Fi/LTE-U Coexistence," 2016 IEEE Global Communications Conference (GLOBECOM), Washington, DC, USA, 2016, pp. 1-6, doi: 10.1109/GLOCOM.2016.7841852.

[12] M. Iqbal, C. Rochman, V. Sathya and M. Ghosh, "Impact of changing energy detection thresholds on fair coexistence of Wi-Fi and LTE in the unlicensed spectrum," 2017 Wireless Telecommunications Symposium (WTS), Chicago, IL, USA, 2017, pp. 1-9, doi: 10.1109/WTS.2017.7943527.

[13] Nsnam „ns-3 Network Simulator." https://www.nsnam.org/, Data dostępu: 22.05.2023r.

ANALIZA MOŻLIWOŚCI ŚWIADCZENIA USŁUG Z ODPOWIEDNIĄ JAKOŚCIĄ W WIELOŁĄCZOWYCH SIECIACH STANDARDU IEEE 802.11BE

ANALYSIS OF QOS PROVISIONING IN IEEE 802.11BE MULTI-LINK NETWORKS

Mikołaj Gwiazdowicz[1]; Marek Natkaniec[1]

[1] Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Kraków,
{mikolaj, natkanie}@agh.edu.pl

DOI: 10.15199/59.2023.4.35

Streszczenie: Autorzy artykułu przedstawiają analizę świadczenia usług z odpowiednią jakością (QoS) w wielołączowych sieciach standardu IEEE 802.11be. Praca koncentrując się na kontrolerze łączy, który wykorzystuje m.in. intensywność zmiany parametru okna rywalizacji mechanizmu EDCA do przewidywania potrzeby nadania priorytetu ruchowi wideo. Proponowany kontroler ma na celu usprawnienie zapewniania QoS poprzez dynamiczne zarządzanie aktywnymi interfejsami.

Abstract: The authors of this paper present an analysis of the quality of service (QoS) provisioning in multi-link IEEE 802.11be networks. The paper focuses on a link controller that uses, among other things, the intensity of change of the contention window parameter of the EDCA mechanism to predict the prioritisation of video traffic. The proposed controller addresses the improvement of QoS provisioning through dynamic management of active interfaces.

Słowa kluczowe: IEEE 802.11be, jakość obsługi, sieci bezprzewodowe, sieci wielołączowe, zarządzanie łączami

Keywords: IEEE 802.11be, link management, multi-link networks, quality of service, wireless networks

1. WSTĘP

Wraz ze stale rosnącym zapotrzebowaniem na wysokiej jakości treści multimedialne, efektywne zarządzanie świadczeniem usług z odpowiednią jakością QoS (*ang. Quality of Service*) w sieciach bezprzewodowych stało się kluczowe. Odpowiedzią na te potrzeby jest ciągły rozwój standardów i mechanizmów wspomagających QoS. Najnowszy standard dotyczący sieci bezprzewodowych IEEE 802.11be [5] wprowadza kolejne usprawnienia mające na celu poprawę m.in. obsługi ruchu wrażliwego na opóźnienia. Rozszerzona do 320 MHz zostanie maksymalna szerokość kanału, zwiększona do 4096-QAM będzie maksymalna modulacja, rozszerzona będzie funkcjonalność MIMO, ale jedną z najważniejszych zmian będzie wprowadzenie transmisji wielołączowej (*ang. Multi-Link Operation*, MLO) [8].

MLO w sieciach WLAN odnosi się do możliwości jednoczesnego wykorzystania wielu interfejsów lub łączy w celu poprawy wydajności sieci, zwiększenia przepustowości lub niezawodności transmisji. Umożliwia urządzeniom typu MLD (*ang. Multi-Link Device*) dystrybucję obciążenia sieci między różne łącza, optymalizując wykorzystanie zasobów i zapobiegając przeciążeniom na pojedynczym łączu. W przypadku prezentowanych wyników badań funkcjonalność ta została zastosowana w celu priorytetyzacji ruchu wideo. Sposób priorytetyzacji został zrealizowany na podstawie przeprowadzonej analizy ruchu w sieciach z urządzeniami typu MLD oraz SLD, czyli klasycznymi urządzeniami stosującymi jednocześnie tylko jedno łącze.

Do najważniejszych aspektów i wniosków tego artykułu można zaliczyć:

1. Przedstawienie nowatorskiego algorytmu dynamicznie zarządzającego priorytetyzacją ruchu dla urządzeń typu MLD.
2. Przedstawienie nowego parametru reprezentującego znormalizowaną intensywność zmiany parametru okna współzawodnictwa CW (*ang. Contention Window*) opisującego stan sieci.
3. Zastosowanie i analizę wydajności mechaniki wielołączowej z wykorzystaniem sieciowego symulatora zdarzeń dyskretnych ns-3 [1].
4. Zastosowanie i analizę działania mechanizmu odwzorowania typu ruchu na łącza (*ang. TID-to-link*) i kontrolera dynamicznie zarządzającego definicją odwzorowania TID-to-link i aktywacją łączy.

Pozostała część artykułu jest zorganizowana w następujący sposób. Rozdział 2 prezentuje przegląd literatury związanej z przeprowadzonymi badaniami. Rozdział 3 opisuje algorytm kontrolera zarządzający priorytetyzacją łączy. Rozdział 4 przedstawia wyniki eksperymentów symulacyjnych proponowanego rozwiązania. Finalnie, rozdziały 5 i 6 zawierają odpowiednio wnioski i ogólne podsumowanie artykułu.

[1] ns-3 sieciowy symulator zdarzeń dyskretnych https://www.nsnam.org

2. PRZEGLĄD LITERATURY

Sieci standardu IEEE 802.11 są bardzo przystępnym środowiskiem oferującym dużą dostępność i łatwość zastosowania dzięki wykorzystaniu kanału radiowego. Jednak właśnie ten aspekt sprawia, że świadczenie usług z odpowiednią jakością jest nietrywialnym zagadnieniem, szczególnie w przypadku ruchu wrażliwego na opóźnienia. Standard IEEE 802.11be, jako kolejny po IEEE 802.11ax, wprowadza szereg usprawnień [4, 9] mających na celu dalsze rozszerzenie możliwości oferowanych przez sieci Wi-Fi, jednocześnie adresując problem aplikacji wrażliwych na opóźnienia. Nowy standard zapewnia nowe narzędzia, jednak nie prezentuje poprawnego sposobu ich stosowania. W pracy przedstawiono innowacyjne rozwiązanie dla mechaniki MLO.

Tematyka mechaniki MLO jest stopniowo poruszana w literaturze [3], lecz z racji wczesnego etapu rozwoju, jej analizy i modyfikacje wykonywane są na podstawie samych obliczeń [9, 1] lub prywatnych symulatorów numerycznych [7]. Tego typu podejście pozwala jedynie na wstępne przebadanie niektórych zagadnień i nie uwzględnia szerszego spektrum zdarzeń i zjawisk, do których dochodzi w sieciach bezprzewodowych.

W sieciach WiFi stacje współzawodniczą ze sobą o dostęp do kanału radiowego na podstawie m.in. parametru CW. Prace naukowe powiązane z parametrem okna współzawodnictwa rozważają kwestie optymalnego doboru samego parametru [2] lub wykrywania niepożądanych zachowań [6]. W przypadku prezentowanych badań, parametr CW posłużył jako opis stanu sieci i metryka dla algorytmu priorytetyzacji ruchu.

3. ALGORYTM PRIORYTETYZACJI

Algorytmu priorytetyzacji ruchu dla urządzeń dysponujących wieloma łączami ma z założenia aktywować dodatkowe łącze w przypadku wykrycia niesprzyjających warunków transmisji na łączu podstawowym, a w przypadku wystarczającej poprawy warunków na łączu podstawowym powinien dezaktywować łącze dodatkowe. Tego typu mechanizm dopasowany został do urządzeń o zasilaniu bateryjnym wymagających transmisji m.in. ruchu wrażliwego na opóźnienia, np. zestawów VR, czy AR.

Negocjacja sposobu wykorzystania łącza dodatkowego jako łącza priorytetowego może być uzgadniana między sąsiadującymi sieciami w ramach kolejnego z mechanizmów IEEE 802.11be, czyli koordynacją między sąsiadującymi punktami dostępowymi.

3.1 Analiza ruchu sieciowego

W celu określenia momentu w którym urządzenie miałoby aktywować i dezaktywować dodatkowe łącze, należało ustalić odpowiednią zależność. Rozważano powiązanie statystyk transmitowanego ruchu widocznych przez urządzenie z metrykami i zmiennymi klas ruchu.

Aby znaleźć satysfakcjonującą zależność wymagane było zebranie danych symulacyjnych do analizy. Opracowano dwuwariantowe scenariusze testowe ze zmienną intensywnością ruchu tła. Pierwszy wariant prezentował działanie urządzeń wykorzystujących pojedyncze (podstawowe) łącze, a drugi urządzeń wykorzystujących oba łącza do transmisji priorytetowego ruchu wideo. Parametry wykonanych symulacji przedstawia Tabela 1. Na potrzeby symulacji zestawiono sieć składającą się z punktu dostępowego (AP) typu MLD oraz dwóch stacji - jednej typu MLD i drugiej typu SLD. Między stacjami, a AP generowano ruch tła klasy BE w obu kierunkach. Priorytetowy ruch wideo generowany był tylko miedzy urządzeniami typu MLD.

Tabela 1: Wybrane parametry symulatora ns-3 zastosowane w symulacjach

Parametr	*Wartość*
Częstotliwość łącza podstawowego	5 GHz
Częstotliwość łącza dodatkowego	6 GHz
Szerokość kanału	20 MHz
Interwał ochronny	800 ns
MCS	5
Mechanizm RTS/CTS	nieaktywny

Wynikiem analizy jest opracowany parametr przedstawiający znormalizowaną intensywność zmiany parametru CW. Jego definicję przedstawia Równanie 1. Na parametr wpływają następujące wartości:

1. ccc - liczba zmian parametru CW w interwale,
2. tp - liczba nadanych pakietów w interwale,
3. tsv - wartość skalowania progu.

$$f_{tsv}(ccc, tp) = \begin{cases} \frac{ccc}{tp*1\%} * tsv, & \text{jeśli } tp > 0 \\ 0, & \text{jeśli } tp == 0 \end{cases} \quad (1)$$

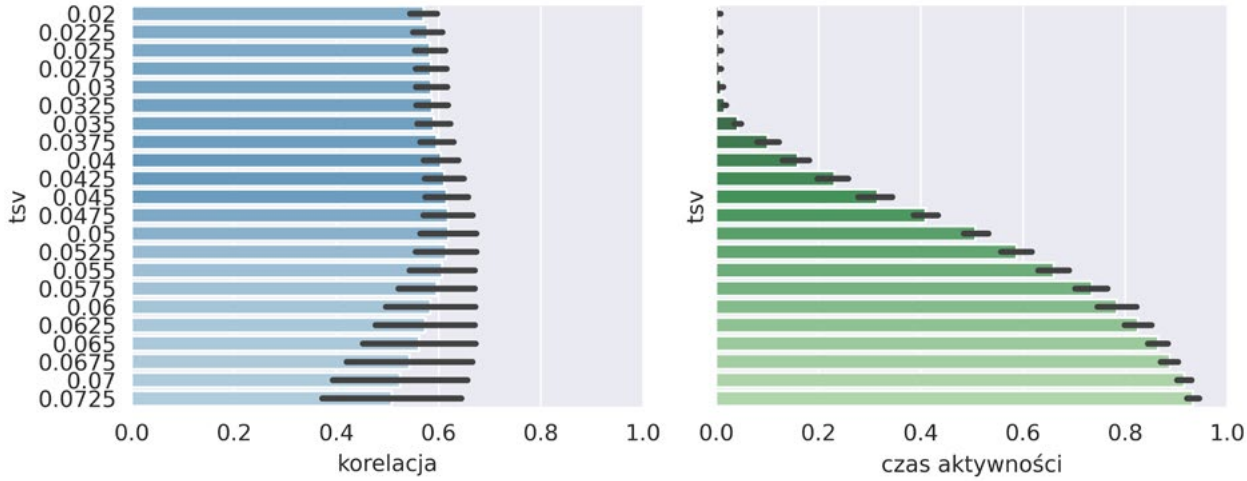
Z założenia wartość parametru od 1.0 i w górę, jest traktowana jako aktywacja łączy priorytetowych, więc w Równaniu 1 można by zastosować funkcję $min(f_{tsv}(ccc, tp), 1.0)$. Jednak na tym etapie taki zabieg mógłby zaburzyć dalszą analizę użyteczności parametru.

Na podstawie zebranych wcześniej danych o parametrach ruchu w sieci wykorzystującej urządzenia z pojedynczym łączem i definicji parametru wykazującego korelację z metryką średniego opóźnienia, możliwe było przedstawienie teoretycznego średniego wykorzystania łącza priorytetowego (dodatkowego). Zarówno poziom korelacji, jak i zależność średniego wykorzystania łącza priorytetowego przedstawia Rysunek 1.

3.2 Opis działania

Zarządzanie stanem łączy realizowane jest przy pomocy kontrolera będącego częścią urządzenia. Kontroler zlicza transmitowane pakiety i zmiany parametru CW dla wybranej klasy ruchu w interwałach czasowych. Na potrzeby implementacji przyjęto okres 100 ms, jako domyślny interwał czasowy. Na początku swojego działania kontroler definiuje odpowiednie zmienne, a następnie wykonuje analizę, której celem jest ustalenie stanu łączy. Analiza wykonywana jest pod koniec każdego interwału, a jej logikę przedstawia Algorytm 1. Działanie kontrolera modyfikują następujące stałe:

1. tsv - wartość skalowania progu wykorzystywana w trakcie podejmowania decyzji o zmianie stanu łączy,
2. cf - współczynnik czasu odświeżania, w trakcie którego nie następuje zmiana stanu łączy,



Rysunek 1: Zależność parametru threshold scaling value (tsv) od korelacji ze średnim opóźnieniem pakietów dla ruchu wideo (po lewej) oraz od czasu aktywności stanu priorytetyzacji dla ruchu wideo (po prawej). Analiza przedstawia wartości teoretyczne uzyskane na podstawie symulacji urządzeń z pojedynczym łączem. Wykres uwzględnia odchylenie standardowe zebranych danych.

3. ws - szerokość okna analizy historii,
4. dv - współczynnik dezaktywacji będący progiem poniżej którego łącza priorytetowe są wyłączane.

Algorytm 1: Logika zarządzania stanem łączy dla ruchu priorytetowego

Stałe: $tsv > 0; cf \geq ws \geq 1; dv > 0$
Zmienne: $history \leftarrow Vector(size : ws; values : 0)$
$lastChangeCounter \leftarrow 0$
$priorityActive \leftarrow false$

```
nccr ← GetCurrentNormalizedCwChangeRate()
history.pushFront(nccr)
lastChangeCounter ← lastChangeCounter + 1
if lastChangeCounter >= cf then
    if priorityActive then
        mean ← sum(history)/ws
        deactivatePriority ← mean < dv
        if deactivatePriority then
            DeactivatePiorityLinks()
            priorityActive ← false
            lastChangeCounter ← 0
        end if
    else
        activatePriority ← nccr ≥ 1.0
        if activatePriority then
            ActivatePiorityLinks()
            priorityActive ← true
            lastChangeCounter ← 0
        end if
    end if
end if
```

4. SYMULACJA ROZWIĄZANIA

W celu przetestowania zaproponowanego rozwiązania rozszerzono symulator ns-3 o mechanizm odwzorowania typu ruchu na łącza (*ang. TID-to-link mapping*) i implementacje kontrolera. Zdefiniowano trzy scenariusze testowe:

1. Ruch kierowany wyłącznie łączem podstawowym.
2. Ruch wideo kierowany łączem podstawowym i dodatkowym; pozostały ruch kierowany łączem podstawowym.
3. Ruch wideo kierowany zgodnie z kontrolerem; pozostały ruch kierowany łączem podstawowym.

W scenariuszu 3 uwzględniono zmienne parametry kontrolera zgodnie z Tabelą 2. Dla uproszczenia parametry cf i ws przyjmują tę samą wartość.

Tabela 2: Parametry symulacyjne kontrolera

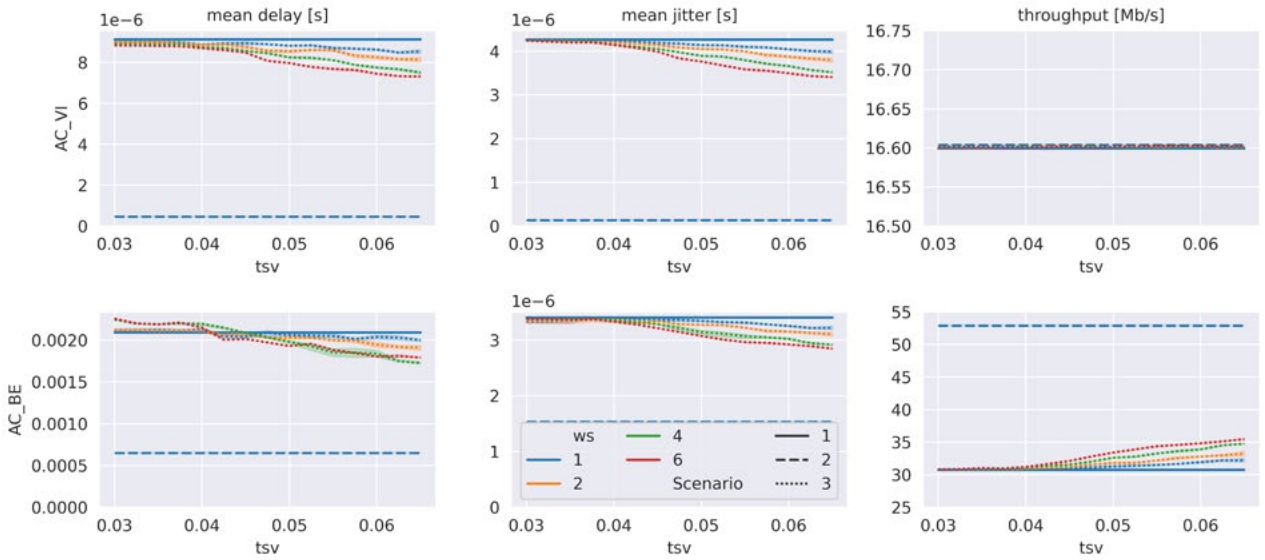
Parametr	*Wartości*
tsv	0.03 : 0.0025 : 0.065
window size	1, 2, 4, 6
deactivation value	0.1 : 0.05 : 0.25

Efekty działania kontrolera przedstawia Rysunek 2. Dane zostały zebrane z 10 przebiegów dla każdej z konfiguracji. Z racji niewielkiego wpływu parametru dv na wyniki, przebiegi te zostały zagregowane.

5. WNIOSKI

Zaproponowany mechanizm kontrolera dynamicznie zarządzającego aktywacją i stanem łączy wykorzystujący znormalizowaną intensywność zmiany parametru CW wykazuje skuteczność w priorytetyzacji ruchu. Kontroler poprawnie reaguje na pogorszenie warunków transmisji ruchu wideo i odpowiednio wykorzystuje możliwości wielołączowych sieci IEEE 802.11be. Zaproponowany parametr tsv pozwala na odpowiednie ustawienie momentu aktywacji priorytetyzacji. Przeprowadzone badania obrazują potencjalne korzyści płynące z proponowanego rozwiązania, takie jak mniejsze opóźnienia, czy wyższa przepustowość w zależności od zapotrzebowania ruchu, a także efektywne zarządzanie energią.

Dalszych badań i dopracowania wymaga warunek dezaktywacji łączy. W aktualnej formie i przy przeana-



Rysunek 2: Wpływ konfiguracji kontrolera (ustawienie parametrów tsv i ws) na wybrane metryki QoS z uwzględnieniem klas ruchu. Metryki QoS idąc od lewej kolumny: średnie opóźnienie [s], średni jitter [s], przepustowość [Mb/s]. Klasy ruchu idąc rzędami od góry: video, best effort.

lizowanym zakresie zmiennej *dv* widoczny jest potencjał do udoskonalenia rozwiązania.

W przypadku transmisji ruchu wideo zarówno łączem dodatkowym, jak i łączem podstawowym, średnia wartość parametru CW w czasie dla łącza podstawowego jest co najmniej dwukrotnie większa, niż w przypadku transmisji tylko łączem podstawowym. Odwrotna zależność ma miejsce w przypadku klasy ruchu BE, gdzie średnia wartość parametru CW w czasie dla łącza podstawowego jest mniejsza.

Zależność znormalizowanej intensywności zmian parametru CW od średniego opóźnienia pozwoliła na implementację kontrolera przedstawionego w tej pracy, jednak podczas badań jeszcze jedna zależność zwracała na siebie uwagę. Dotyczy to zależności częstotliwości zmian parametru CW od średniego jittera, która charakteryzowała się znacznie wyższym współczynnikiem korelacji, jednak jej zastosowanie wykraczało poza przeprowadzane prace.

6. PODSUMOWANIE

Niniejszy artykuł przedstawia analizę świadczenia usług z odpowiednią jakością w wielołączowych sieciach standardu IEEE 802.11be. Przedstawiono w nim oryginalny mechanizm kontrolera łączy, który wykorzystuje m.in. zależność zmian parametru okna rywalizacji do świadczenia priorytetyzacji dla ruchu wideo. Przeprowadzone badania wykazały, że zaproponowany kontroler skutecznie poprawia metryki QoS w zależności od zapotrzebowania ruchu.

LITERATURA

[1] Naik Gaurang, Dennis Ogbe, and Jung-Min Jerry Park, 2021. "Can Wi-Fi 7 Support Real-Time Applications? On the Impact of Multi Link Aggregation on Latency", ICC 2021 - IEEE International Conference on Communications, Montreal, Canada, pp. 1-6.

[2] Wydmański Witold, and Szymon Szott, 2021. "Contention Window Optimization in IEEE 802.11ax Networks with Deep Reinforcement Learning", 2021 IEEE Wireless Communications and Networking Conference (WCNC), Nanjing, China, pp. 1-6.

[3] López-Raventós Álvaro, and Boris Bellalta, 2022. "Dynamic Traffic Allocation in IEEE 802.11be Multi-Link WLANs", IEEE Wireless Communications Letters, vol. 11 (7): 1404-1408.

[4] Lopez-Perez David, Adrian Garcia-Rodriguez, Lorenzo Galati-Giordano, Mika Kasslin, and Klaus Doppler, 2019. "IEEE 802.11be Extremely High Throughput: The Next Generation of Wi-Fi Technology Beyond 802.11ax", IEEE Communications Magazine, vol. 57 (9): 113-119.

[5] 2023. "IEEE Draft Standard for Information technology–Telecommunications and information exchange between systems Local and metropolitan area networks–Specific requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment: Enhancements for Extremely High Throughput (EHT)" IEEE P802.11be/D3.0, pp.1-999.

[6] Szott Szymon, Marek Natkaniec, and Albert Banchs, 2009. "Impact of Misbehaviour on QoS in Wireless Mesh Networks", NETWORKING 2009, Berlin, Heidelberg, pp. 639-650.

[7] Huang Kaiwen, Lei Huang, Yingqiao Quan, Haoyuan Du, Chaoming Luo, Liuming Lu, and Ronghui Hou, 2022. "Mutli-Link Channel Access Schemes for IEEE 802.11be Extremely High Throughput", IEEE Communications Standards Magazine, vol. 6 (3): 46-51.

[8] López-Raventós Álvaro, and Boris Bellalta, 2022. "Multi-Link Operation in IEEE 802.11be WLANs," IEEE Wireless Communications, vol. 29 (4): 94-100.

[9] Adame, Toni, Marc Carrascosa-Zamacois, and Boris Bellalta, 2021. "Time-Sensitive Networking in IEEE 802.11be: On the Way to Low-Latency WiFi 7" Sensors 21, no. 15: 4954.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

ZASTOSOWANIE ALGORYTMÓW WIELORĘKICH BANDYTÓW DO WYBORU SZYBKOŚCI TRANSMISJI W SIECIACH IEEE 802.11AX

APPLYING MULTI-ARMED BANDIT ALGORITHMS FOR DATA RATE SELECTION IN IEEE 802.11AX NETWORKS

Maksymilian Wojnar; Wojciech Ciężobka

AGH Akademia Górniczo-Hutnicza, Kraków, {imię.nazwisko}@agh.edu.pl

DOI: 10.15199/59.2023.4.36

Streszczenie: Wybór szybkości transmisji jest kluczowym problemem sieci IEEE 802.11. Można traktować go jako zagadnienie wielorękich bandytów. W tej pracy zaimplementowano algorytmy wielorękich bandytów używając autorskiej biblioteki programistycznej. Przeprowadzono studium parametrów oraz testy działania rozwiązań w symulatorze ns-3. Wyniki pokazują, że zastosowanie omawianych algorytmów może znacząco poprawić wydajność gęstych sieci Wi-Fi.

Abstract: Data rate selection is a typical problem of Wi-Fi networks. We present this problem in IEEE 802.11 networks as a multi-armed bandit (MAB) problem. We implement MAB algorithms using our library for rapid prototyping of reinforcement learning algorithms. We conduct a parameter study and test the performance of the solutions in the ns-3 simulator. Our results show that using MAB algorithms can significantly improve the performance of dense Wi-Fi networks.

Słowa kluczowe: 802.11, uczenie maszynowe, wieloręcy bandyci, wybór szybkości transmisji

Keywords: 802.11, data rate selection, machine learning, multi-armed bandits

1. WSTĘP

Aby osiągnąć wysoką przepustowość w sieciach IEEE 802.11 (Wi-Fi), należy dynamicznie wybierać odpowiednią szybkość transmisji (schemat modulacji i kodowania, ang. *modulation and coding scheme*, MCS). Zbyt wysoki MCS powoduje konieczność retransmisji, a zbyt niski – nieefektywne wykorzystanie kanału. Wyboru MCS dokonuje menedżer szybkości transmisji. Typowe menedżery, takie jak Minstrel [2], często charakteryzują się słabą wydajnością w gęstych sieciach, gdyż mylą kolizje między ramkami z ustawieniem zbyt wysokiego MCS.

We współczesnych sieciach Wi-Fi popularne staje się używanie uczenia ze wzmocnieniem (ang. *reinforcement learning*, RL), jako sposobu na poprawę wydajności sieci [5]. Ogólny schemat RL jest następujący: agent (np. menedżer szybkości transmisji) obserwuje stan środowiska (np. liczbę udanych i nieudanych transmisji), wybiera akcję (np. MCS dla następnej transmisji) na podstawie stanu środowiska i otrzymuje nagrodę (np. chwilową przepustowość) po podjęciu tej akcji. Szczególnym przypadkiem RL jest problem wielorękich bandytów (ang. *multi-armed bandit*, MAB). Jest to klasa problemów, w których występuje kompromis pomiędzy eksploracją (próbą zdobycia nowej wiedzy) oraz eksploatacją (wykorzystaniem zdobytej wiedzy do podejmowania decyzji). Agent ma ograniczone zasoby: może podejmować decyzje tylko skończoną liczbę razy, zatem powinien jak najszybciej odkryć akcję dającą najwyższą nagrodę. Nagroda za każde ramię (wybór pewnej akcji) może się zmieniać w czasie, więc problemy MAB mogą być również niestacjonarne. Problem wyboru odpowiedniej szybkości transmisji można traktować jako niestacjonarny kontekstowy problem wielorękiego bandyty. Kontekstowy bandyta rozszerza model o podejmowanie decyzji zarówno w oparciu o stan środowiska, jak i charakterystykę ramion bandyty. W przypadku wyboru szybkości transmisji, zbiorem akcji jest zbiór możliwych wartości MCS, natomiast kontekstem może być szybkość transmisji danych dla danego MCS.

Celem badań jest pokazanie, że zastosowanie algorytmów MAB do wyboru szybkości transmisji w sieciach IEEE 802.11ax prowadzi do poprawy wydajności względem dotychczas stosowanych algorytmów. W literaturze rozważano algorytmy MAB do wyboru szybkości transmisji [3]. Nowością referatu jest analiza większej liczby algorytmów w symulatorze ns-3 oraz implementacja autorskiej biblioteki programistycznej. Rozdział 2 zawiera opis zastosowanych narzędzi oraz wybrane algorytmy MAB. Rozdział 3 zawiera opis przeprowadzonego studium parametrów wybranych algorytmów, natomiast rozdział 4 poświęcony jest ewaluacji proponowanych rozwiązań w symulatorze ns-3. Pracę podsumowano w rozdziale 5.

2. ALGORYTMY I NARZĘDZIA

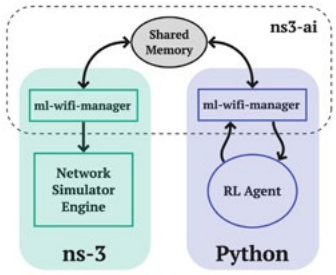
Wybrano kilka powszechnie znanych algorytmów MAB, aby zbadać ich wydajność jako menedżerów wyboru szybkości transmisji. Ze względu na obszerność dokładnego opisu wszystkich zastosowanych

algorytmów, w tej pracy zamieszczamy jedynie krótki opis ich działania wraz z podsumowaniem kluczowych parametrów. Opis większości algorytmów można znaleźć m.in. w [4].

Pierwszy algorytm, ε-greedy, to podstawowe rozwiązanie kompromisu pomiędzy eksploracją, a eksploatacją. Agent z prawdopodobieństwem ε podejmuje losową akcję, w przeciwnym przypadku wybiera obecnie najlepszą akcję. Parametr α steruje szybkością zaniku estymowanej wartości akcji, natomiast ν to wartość początkowa estymaty. Upper Confidence Bound (UCB) rozszerza ε-greedy o człon z licznikiem prób danej akcji, którego wagę określa parametr c. Tak prosta modyfikacja pozwala na asymptotycznie optymalne działanie agenta w stacjonarnym środowisku. Wartość γ steruje szybkością zaniku wykorzystywanego w niestacjonarnym przypadku. Algorytm EXP3 [1] przypisuje akcjom wagi, na podstawie których wyznaczane jest prawdopodobieństwo podjęcia danej akcji. Wagi z każdym krokiem są wykładniczo redukowane. Parametr γ określa stopień eksploracji algorytmu oraz szybkość wykładniczego zaniku wag. W Softmax agent uczy się preferencji nad dyskretnym zbiorem akcji wykorzystując gradient funkcji z szybkością określoną przez stałą uczącą lr. Polityka agenta to funkcja *softmax* nad nauczonymi preferencjami z temperaturą τ. Parametr α steruje szybkością zaniku preferencji. Natomiast w Thompson Sampling (TS) [3] agent podejmuje akcję w oparciu o wewnętrzny, probabilistyczny model. W naszej implementacji prawdopodobieństwo sukcesu modelujemy rozkładami beta, których parametry ulegają wykładniczemu zanikowi z szybkością w.

Wymienione algorytmy zaimplementowano w autorskiej bibliotece *Reinforced-lib*[1], zbudowanej przy użyciu podejścia modułowego. Biblioteka została stworzona do szybkiego prototypowania rozwiązań RL, ze szczególnym uwzględnieniem sieci bezprzewodowych. Dostarcza ona narzędzi do oceny i analizy zaimplementowanych metod. Agentów zaimplementowano w JAX – nowoczesnym frameworku stosowanym w dziedzinie uczenia maszynowego. Biblioteka posiada obszerną dokumentację[2], jest udostępniona na otwartej licencji oraz umieszczona w publicznym repozytorium[3].

Aby zbadać wydajność sieci IEEE 802.11, zastosowano symulator ns-3[4]. Do integracji autorskiej biblioteki z symulatorem, użyto modułu ns3-ai [6], który zapewnia komunikację pomiędzy procesami ns-3 i Pythona przez pamięć współdzieloną. Rysunek 1 przedstawia proces wyboru MCS. Najpierw symulator pyta moduł *ml-wifi-manager* o MCS kolejnej transmisji i moduł po stronie C++ przekazuje odpowiednie dane do procesu Pythona. Stan sieci trafia do agenta – algorytmu zaimplementowanego w bibliotece. Agent podejmuje decyzję, a informacja jest przekazywana z powrotem do symulatora i ramka jest wysyłana z użyciem wyznaczonego MCS.



Rysunek 1: Schemat komunikacji pomiędzy biblioteką *Reinforced-lib*, a symulatorem ns-3.

Ponieważ obliczenia w ns-3 są czasochłonne, do znalezienia najlepszych parametrów dla badanych algorytmów stworzono *ra-sim*[5] – prosty symulator wyboru szybkości transmisji w sieciach Wi-Fi. Obsługuje on podstawową mobilność, wiele transmitujących stacji oraz agregację A-MPDU (ang. *Aggregated MAC Protocol Data Unit*). Kanał komunikacyjny modelowany jest przez dwa równania wyprowadzone empirycznie poprzez porównanie z wynikami uzyskanymi z symulacji w ns-3:

$$P(s=1|\gamma,\mu) = \Phi\left(2\sqrt{2}\,(\gamma-\gamma_\mu)\right), \quad (1)$$

$$P(c=1|N^{STA}) = 0.15489\cdot\ln\left(1.0310\cdot N^{STA}\right), \quad (2)$$

gdzie s wskazuje, czy transmisja zakończyła się sukcesem, γ oznacza aktualny stosunek siły sygnału do szumu (ang. *signal to noise ratio*, SNR), μ to MCS użyty do transmisji, Φ to dystrybuanta standardowego rozkładu normalnego, γ_μ oznacza minimalny SNR wymagany do transmisji z danym MCS, c wskazuje, czy wystąpiła kolizja, a N^{STA} to liczba jednocześnie nadających stacji. Równanie (1) opisuje prawdopodobieństwo udanej transmisji, natomiast (2) – prawdopodobieństwo kolizji.

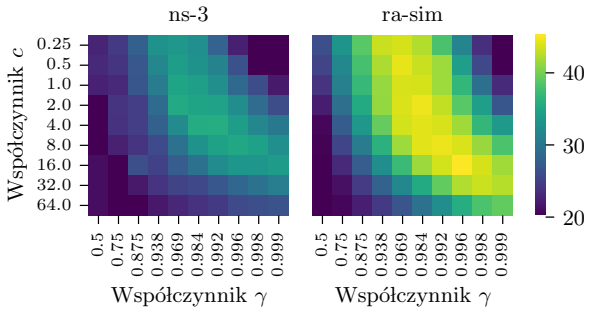
3. STUDIUM PARAMETRÓW

Do studium parametrów użyto *ra-sim*, który znacząco (nawet kilkunastokrotnie) przyspiesza symulacje. Aby zweryfikować równoważność badania parametrów za pomocą *ra-sim* i ns-3, przeprowadzono poszukiwanie parametrów UCB w obydwu symulatorach. Scenariusz porównawczy zakłada grupę 10 stacji oddalających się od punktu dostępowego (ang. *access point*, AP) ze stałą prędkością 1 m/s przez 50 s. Parametry symulacji umieszczone są w tabeli 1. Studium przeprowadzono w scenariuszu dynamicznym, aby przetestować agentów w bardziej realistycznym, niestacjonarnym przypadku. Wybrano dodatkowo większą liczbę stacji, aby utrudnić zadanie ze względu na występujące kolizje.

[1] https://github.com/m-wojnar/reinforced-lib/
[2] https://reinforced-lib.readthedocs.io/
[3] https://pypi.org/project/reinforced-lib/
[4] https://www.nsnam.org/
[5] https://github.com/m-wojnar/reinforced-lib/tree/main/examples/ra-sim

Tabela 1: Parametry symulacji

Parametr	*Wartość*
PHY/MAC	IEEE 802.11ax
Pasmo	5 GHz
Strumienie przestrzenne	1
Szerokość kanału	20 MHz
Okres ochronny	3200 ns
Model kanału	log-distance
Agregacja ramek	A-MPDU
Wielkość ramki	1500 B
Protokół RTS/CTS	Wyłączony
Ruch generowany przez stację	Uplink, 125 Mb/s



Rysunek 2: Porównanie wyników z symulatorów ns-3 oraz *ra-sim* na przykładzie agenta UCB.

Wykonano przeszukiwanie siatki w przestrzeni możliwych wartości parametrów. Każdą kombinację parametrów pięciokrotnie oceniono mierząc średnią przepustowość sieci. Różnica w przepustowości pomiędzy symulatorami (rysunek 2) wynika z uproszczeń zastosowanych w *ra-sim* (nie uwzględnia on wszystkich reguł dostępu do kanału z IEEE 802.11, np. nie podwaja okna współzawodnictwa po kolizji). Najważniejszym aspektem porównania jest jednak konfrontacja kształtów przestrzeni otrzymanych z symulatorów – są one zbliżone, w szczególności ekstremum znajduje się w podobnym punkcie. Stąd wniosek, że zastosowane podejście do badania parametrów jest wystarczające. Najlepsze kombinacje parametrów zawiera tabela 2, natomiast wszystkie dane oraz wykresy zostały opublikowane na platformie *Zenodo*[6].

4. OCENA WYDAJNOŚCI

Znając optymalny zestaw parametrów (tabela 2), sprawdzono działanie algorytmów w symulatorze ns-3. Aby zagwarantować zróżnicowaną ocenę, wprowadzono dwa rodzaje scenariuszy z różną dynamiką stacji. Wyniki porównano z menedżerami zaimplementowanymi w ns-3, tj. Minstrel (jako popularny bazowy algorytm) oraz Ideal (jako górna granica). Parametry symulacji umieszczone są w tabeli 1. Udostępniono kod symulacji na otwartej licencji[7].

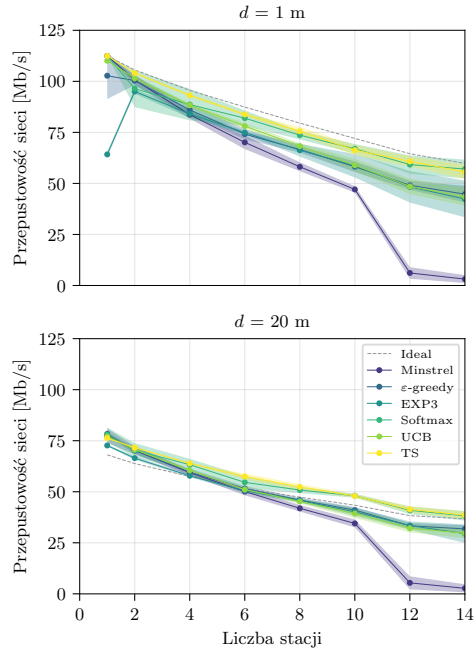
Tabela 2: Najlepsze parametry zaimplementowanych algorytmów w problemie wyboru szybkości transmisji

Algorytm	*Parametr*	*Wartość*
ε-greedy	ε	0.008
ε-greedy	α	0.032
ε-greedy	ν	64.0
EXP3	γ	0.128
Softmax	lr	2.048
Softmax	α	0.016
Softmax	τ	4.0
Thompson Sampling	w	2.0
Upper Confidence Bound	c	16.0
Upper Confidence Bound	γ	0.996

4.1 Scenariusz stacjonarny

W scenariuszu stacjonarnym grupa nieruchomych stacji, umieszczonych w danej odległości od AP, transmituje w pełnym nasyceniu. Analizujemy stacje w odległości 1 m (scenariusz wyidealizowany, nie występują błędy kanału, a jedynie kolizje) oraz bardziej realistyczny scenariusz z odległością 20 m. Każdy punkt na wykresie to średnia z przynajmniej pięciu niezależnych przebiegów symulacji.

Wyniki pokazują, że TS i Softmax zapewniają dobrą wydajność w obu wariantach. ε-greedy, EXP3 oraz UCB wypadają nieco gorzej (rysunek 3), jednak wszystkie algorytmy MAB wyprzedzają menedżer Minstrel, szczególnie dla gęstych sieci, w których liczba stacji jest większa lub równa 12. Tabela 3 zawiera porównanie średnich przepustowości.

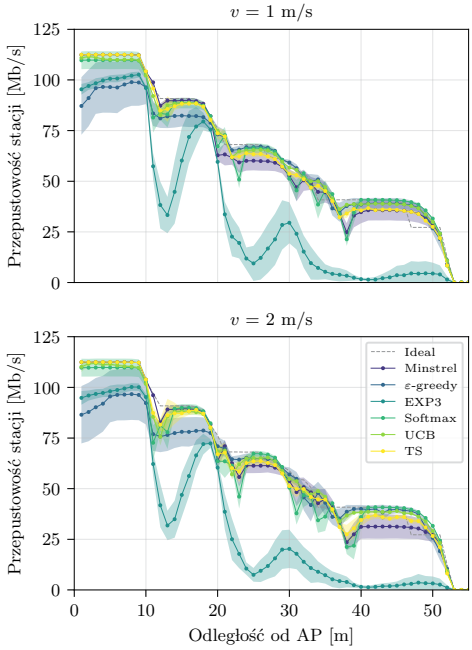


Rysunek 3: Scenariusz stacjonarny. Pasma reprezentują 99-procentowe przedziały ufności.

[6]https://zenodo.org/record/7944539
[7]https://github.com/m-wojnar/reinforced-lib/tree/main/examples/ns-3-ra

Tabela 3: Średnia przepustowość w badanych scenariuszach w Mb/s. Kolumna A – scenariusz stacjonarny w odległości 1 m, B – scenariusz stacjonarny dla 20 m, C – scenariusz mobilny dla prędkości 1 m/s, D – scenariusz mobilny dla 2 m/s.

Algorytm	A	B	C	D
Ideal	84.61	50.87	63.45	63.47
Minstrel	60.17	42.74	61.04	61.18
ε-greedy	72.73	51.49	59.31	57.45
EXP3	66.72	49.55	34.77	32.78
Softmax	79.54	55.54	61.50	60.11
TS	81.47	56.20	61.66	60.64
UCB	74.64	50.74	62.68	61.62



Rysunek 4: Scenariusz mobilny. Pasma reprezentują 99-procentowe przedziały ufności.

4.2 Scenariusz mobilny

Scenariusz mobilny zakłada pojedynczą stację oddalającą się od AP ze stałą prędkością: 1 m/s lub 2 m/s. Każdy punkt na wykresie to średnia z 20 niezależnych przebiegów symulacji. Różnica w przepustowości pomiędzy Minstrel oraz Ideal jest szczególnie mała (tabela 3), jednak UCB udaje się zmieścić pomiędzy nimi (rysunek 4). Softmax oraz TS również mają dobre wyniki, są bardziej wrażliwe na zmiany MCS, ale ich szybka zbieżność pozwala na uzyskanie dobrej przepustowości w ciągu całej symulacji. Wydajność najprostszego ε-greedy poprawia się wraz z odległością. Takie zachowanie może być wynikiem wolnej zbieżności algorytmu oraz strat wynikających ze stałego współczynnika eksploracji. Z kolei wydajność EXP3 jest słaba od początku, algorytm jest wrażliwy na niepowodzenia transmisji i agresywnie zmniejsza wartość MCS, a dodatkowo charakteryzuje się wolną zbieżnością.

Na rysunku 4 dla wszystkich algorytmów widoczne są charakterystyczne zafalowania w miejscach zmiany szybkości transmisji. Są one spowodowane koniecznością adaptacji algorytmów do nowego trybu. Po osiągnięciu pewnej odległości (co wiąże się z pogorszeniem jakości kanału), menedżery szybkości odnotowują straty w transmisji pakietów. Ponieważ czas reakcji algorytmów jest niezerowy (w przeciwnym wypadku byłyby one podatne na losowe krótkotrwałe zmiany siły sygnału), występuje chwilowy spadek przepustowości przed przełączeniem na niższy tryb.

5. WNIOSKI

W artykule przedstawiono zastosowanie algorytmów MAB do wyboru szybkości transmisji w sieciach IEEE 802.11ax. Algorytmy zaimplementowano w bibliotece *Reinforced-lib*, studium parametrów przeprowadzono w symulatorze *ra-sim*, a w ns-3 sprawdzono wydajność menedżerów szybkości transmisji wykorzystujących wybrane algorytmy MAB. Wyniki pokazały, że TS i UCB są w stanie osiągać wysoką wydajność, przewyższając popularnie stosowany menedżer Minstrel oraz zbliżając się do górnej granicy wyznaczonej przez menedżer Ideal.

Zgodnie z naszą wiedzą, badane algorytmy nie były do tej pory implementowane w rzeczywistych sieciach. Mają one jednak duży potencjał: nie wymagają dedykowanych rozwiązań sprzętowych, a ich koszt obliczeniowy jest stosunkowo niski. Podsumowując, zastosowanie algorytmów wielorękich bandytów w wyborze szybkości transmisji w sieciach Wi-Fi może znacząco poprawić wydajność, zwłaszcza w przypadku gęstych sieci.

PODZIĘKOWANIA

Praca powstała w wyniku realizacji projektu badawczego o nr DEC-2020/39/I/ST7/01457 finansowanego ze środków Narodowego Centrum Nauki.

LITERATURA

[1] Auer P. et al. 2002. „The Nonstochastic Multiarmed Bandit Problem”. SIAM J. Comput. 32: 48-77.

[2] Fietkau F. 2010. „Minstrel HT Linux Wireless”. https://lwn.net/Articles/376765/.

[3] Krotov A., A. Kiryanov, E. Khorov. 2020. „Rate Control With Spatial Reuse for Wi-Fi 6 Dense Deployments”. IEEE Access 8: 168898–909.

[4] Sutton R., A. Barto. 2018. „Reinforcement Learning: An Introduction”. MIT Press.

[5] Szott S. et al. 2022. „Wi-Fi Meets ML: A Survey on Improving IEEE 802.11 Performance With Machine Learning”. IEEE Commun. Surveys & Tuts. 24(3): 1843–93.

[6] Yin H. et al. 2020. „Ns3-Ai: Fostering Artificial Intelligence Algorithms for Networking Research”. 2020 Workshop on ns-3, 57–64.

ANALIZA DZIAŁANIA MECHANIZMÓW POTWIERDZEŃ RAMEK DLA TRYBU MU OFDMA SIECI STANDARDU IEEE 802.11AX

AN ANALYSIS OF ACKNOWLEDGMENT MECHANISMS FOR THE MU OFDMA TRANSMISSION OF THE IEEE 802.11AX NETWORKS

Marek Natkaniec[1]; Michał Kaszuba[2]

[1] AGH Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Kraków, natkanie@agh.edu.pl
[2] AGH Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Kraków, kaszubam@student.agh.edu.pl

DOI: 10.15199/59.2023.4.37

Streszczenie**: Standard IEEE 802.11ax dzięki użyciu techniki OFDMA umożliwia realizację równoczesnej transmisji pomiędzy punktem dostępu i wieloma stacjami. Praca zawiera analizę trzech różnych mechanizmów potwierdzeń ramek dla trybu MU OFDMA. Przeanalizowane zostały dwa scenariusze, w których zmienna była liczba stacji oraz wielkość pakietu. Wyniki zostały porównane z mechanizmem potwierdzeń stosowanym dla techniki OFDM. Uzyskane wyniki pokazują, który z mechanizmów powinien być użyty dla określonej topologii sieciowej, aby zapewnić odpowiednią wydajność pracy sieci bezprzewodowej.**

Abstract**: The IEEE 802.11ax standard, using the OFDMA technique, enables simultaneous transmission between access point and multiple stations. The paper includes an analysis of three different frame acknowledgment mechanisms for the MU OFDMA mode. Two scenarios were examined, where the variable was the number of stations and the packet length. The results were compared with the acknowledgment mechanism used for the OFDM technique. The obtained results indicate which mechanism should be used for a given network topology to ensure appropriate wireless network performance.**

Słowa kluczowe**: IEEE 802.11ax, mechanizmy potwierdzeń ramek, OFDMA, transmisja wieloużytkownikowa.**

Keywords**: IEEE 802.11ax, frames acknowledgement mechanisms, OFDMA, multiuser transmission.**

1. WSTĘP

Internet Rzeczy oraz ciągle rosnąca liczba użytkowników mobilnych stawia coraz większe wyzwania lokalnym sieciom bezprzewodowym WLANs (*Wireless Local Area Networks*). Wykorzystywane przez sieci WLAN pasma częstotliwości 2.4 GHz oraz 5 GHz mają ograniczone zasoby, a użytkownicy wymagają, aby połączenie z Internetem było szybkie i niezawodne. Z pomocą przychodzą najnowsze rozszerzenia standardu sieci IEEE 802.11 [1] opisujące szereg mechanizmów poprawiających jakość pracy sieci bezprzewodowych.

Najnowszym dostępnym komercyjnie standardem jest obecnie IEEE 802.11ax, który poprawia równoległą transmisję od punktu dostępu do wielu użytkowników za pomocą techniki OFDMA (*Orthogonal Frequency Division Multiple Access*). IEEE 802.11ax został wprowadzony w 2020 roku i umożliwia transmisję danych z szybkością do 9,6 Gbit/s. W artykule zostaną przeprowadzone badania dla dwóch topologii sieciowych, wykorzystujących różne konfiguracje oraz mechanizmy potwierdzeń w trybie transmisji w łączu w dół DL OFDMA w porównaniu ze standardową transmisją OFDM (*Orthogonal Frequency Division Multiplexing*). Do badań zostanie użyty symulator zdarzeniowy ns-3.

Praca składa się z sześciu rozdziałów. Rozdział 2 przedstawia przegląd literatury. W rozdziale 3 przybliżono technikę OFDMA stosowaną w standardzie IEEE 802.11ax, a w kolejnym rozdziale omówiono mechanizmy potwierdzeń ramek. W rozdziale 5 zawarto wyniki badań symulacyjnych przeprowadzonych dla dwóch różnych scenariuszy pracy sieci bezprzewodowej Wi-Fi. Artykuł kończy rozdział 6, który zawiera podsumowanie niniejszej pracy.

2. PRZEGLĄD LITERATURY

2.1. MU OFDMA

Artykuł [2] opisuje analizę wydajności transmisji w górę MU OFDMA. Modelem testowym był jeden punkt dostępowy z n stacjami rywalizującymi o dostęp do kanału radiowego. Badania prowadzone w warunkach nasycenia dowiodły, że przy rosnącej liczbie stacji nie ma znaczącego spadku przepustowości, jednakże pojawiają się kolizje. Przydzielenie większej liczby jednostek RU zmniejsza liczbę kolizji. Zwiększająca się liczba RU ma również bezpośredni wpływ na czas oczekiwania na transmisję, a w konsekwencji na opóźnienie. Przepustowość transmisji w łączu w dół OFDMA dla nienasyconej sieci została przeanalizowana w pracy [3]. Omawiany wyżej wzrost liczby jednostek zasobów wydzielonych do dostępu losowego również pozytywnie wpływa na transmisję DL. Wykazano, że zwiększenie możliwości transmisji UL ma bezpośrednie przełożenie na transmisję w łączu w dół. Autorzy artykułu [4] przedstawili mechanizm maksymalizujący przepustowość dla transmisji MU OFDMA zarówno w łączu w dół jak i w górę. W najbardziej wymagającym przypadku uzyskano zysk o 50%. Poprawie uległy również średnie opóźnienia transmisji. W pracy [5] zaproponowano nowatorski mechanizm transmisji dla wielu użytkowników w sieciach IEEE 802.11ax, który korzysta z techniki OFDMA oraz MU-MIMO. Punkt dostępowy szacuje i gromadzi parametry kanału z ramki OFDMA łącza w górę, a te informacje są następnie wy-

korzystywane do znalezienia optymalnej grupy odbiorników MU-MIMO łącza w dół, aby zmaksymalizować wydajność systemu. Autorzy opracowali nowy algorytm alokacji zasobów i strukturę ramek. Przeprowadzone symulacje wskazują na 4-krotny wzrost przepustowości.

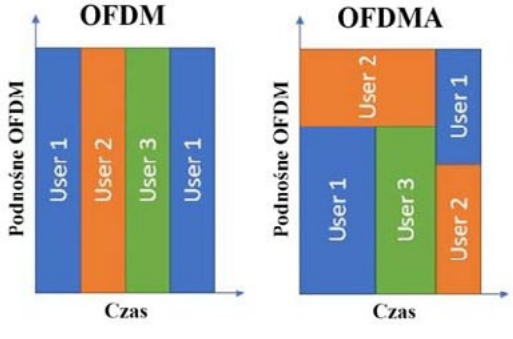
2.2. Potwierdzenia blokowe

Artykuł [6] opisuje badania przeprowadzone w zmiennych warunkach kanału radiowego, dla różnych wartości parametru BER (*Bit Error Rate*). Z transmisji prowadzonej między dwiema stacjami wynika, że potwierdzenie blokowe zwiększa efektywność pracy sieci w dobrych warunkach kanału w przypadku transmisji TCP, a na transmisję UDP nie znaczącego wpływu. Autorzy, w pracy [7] opisali efektywność działania blokowego potwierdzenia dla kanału wolnego od błędów. Udowodnili, że blokowe ACK nie jest przydatne dla niskich szybkości transmisji oraz krótkich czasów TXOP, podczas gdy jest bardzo efektywne dla dużych szybkości transmisji. Wykazali również, że w transmisji opartej na współzawodnictwie liczba stacji nie wpływa na optymalny wybór między normalnym potwierdzeniem, a potwierdzeniem blokowym. Przeprowadzono także analizę efektywności energetycznej i odporności na błędy transmisji MU OFDMA i blokowych potwierdzeń. Potwierdzono, że proponowane mechanizmy i nowe techniki wpływały korzystnie na wydajność pracy sieci.

Niniejsza praca łączy dwie wymienione wyżej techniki – MU OFDMA i stosowanie potwierdzeń blokowych. Według naszej najlepszej wiedzy, podobne badania nie były dotąd opisywane w literaturze.

3. TECHNIKA OFDMA W IEEE 802.11AX

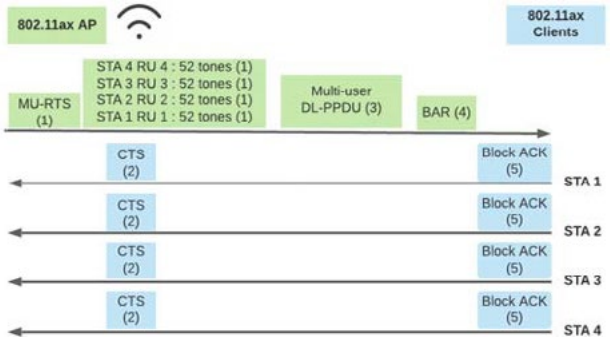
Standard IEEE 802.11ax wykorzystuje technikę OFDMA jako mechanizm warstwy PHY, aby jednocześnie obsługiwać wielu użytkowników [1]. OFDMA jest wielodostępowym wariantem OFDM, w którym następuje przypisanie podzbiorów podnośnych OFDM, zwanych jednostkami zasobów RUs (*Resource Units*) do różnych użytkowników. Umożliwia to równoczesną transmisję ramek danych, kontrolnych i zarządzających z użyciem DL OFDMA do i/lub korzystając z UL OFDMA od kilku użytkowników. Zastosowanie techniki OFDMA zmniejsza nadmiar związany z transmisją w kanale radiowym. IEEE 802.11ax pozwala na obsługę kanałów o szerokości 20/40/80/160 (80+80) MHz w paśmie częstotliwości 5 GHz. Na Rys. 1 można dostrzec, że w przypadku użycia techniki OFDM, w określonym czasie tylko jeden użytkownik zajmuje cały kanał radiowy.



Rys. 1. Porównanie techniki OFDM i OFDMA

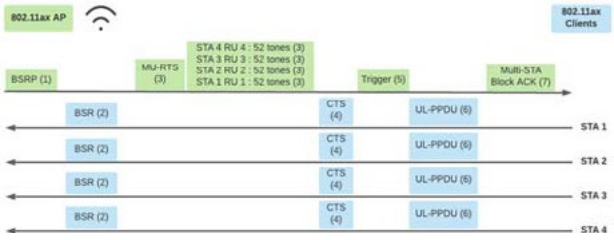
Dla OFDMA punkt dostępowy dzieli kanały na RU, które przydziela odpowiednio użytkownikom. Dzięki temu możliwa jest równoczesna transmisja na przydzielonych podnośnych, bez kolizji między stacjami, co zmniejsza średnie opóźnienie transmisji i jego wariancję.

Rys. 2 przedstawia przykład transmisji dla kierunku DL OFDMA. Transmisja odbywa się, gdy punkt dostępowy ma dane gotowe do wysłania do wielu stacji. W tym celu wysyła on ramkę MU-RTS, na którą stacje odpowiadają ramką CTS. Następnie, na podstawie określonych RU, przydzielane są podnośne do stacji. Po zakończonej równoległej transmisji danych punkt dostępowy wysyła ramkę BAR (*Block Acknowledgement Request*) jako blokowe żądanie potwierdzenia, na którą stacje odpowiadają równocześnie ramką blokowego potwierdzenia.



Rys. 2. Przykład transmisji DL-OFDMA [8]

W przeciwieństwie do OFDM, dla UL/DL OFDMA nie występuje rywalizacja o dostęp do kanału radiowego. Ramka MU-RTS zawiera informacje o przypisanych podnośnych, RU oraz o długości trwania transmisji. Punkt dostępowy decyduje jaki rodzaj transmisji MU (*Multi User*) zainicjować. Przykładową transmisję dla kierunku UL OFDMA przedstawia Rys. 3, gdzie AP rozpoczyna proces wysyłając ramkę BSRP (*Buffer Status Report Poll*), by sprawdzić ile danych w stacjach jest gotowych do wysłania. Klienci odpowiadają ramkami BSR (*Buffer Status Report*), dzięki którym punkt dostępowy może rozdzielić podnośne i jednostki zasobów. Następuje wymiana ramek MU-RTS i CTS tak jak w przypadku transmisji DL. Kluczowym momentem jest wysłanie ramki Trigger, która wyzwala równoczesną transmisję od kilku stacji. Po poprawnym przesłaniu danych punkt dostępowy odpowiada blokowym potwierdzeniem Block ACK.



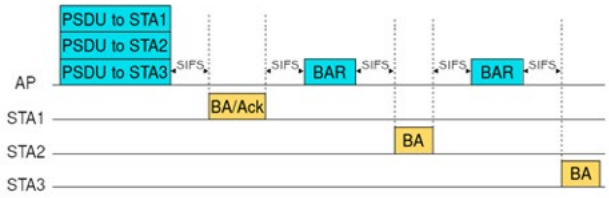
Rys. 3. Przykład transmisji UL-OFDMA [8]

4. MECHANIZMY POTWIERDZEŃ

Blokowe mechanizmy potwierdzeń zostały zaproponowane po raz pierwszy w rozszerzeniu IEEE 802.11e wraz z czasem TXOP (*Transmission Opportunity*). TXOP to okres czasu, w którym stacja może nadawać kolejne

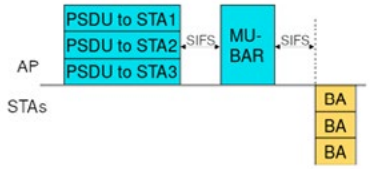
ramki oddzielone czasem SIFS bez konieczności wymiany ramki potwierdzeń. Jeśli transmisja ramek przekroczy limit TXOP, konieczna jest fragmentacja i dokończenie transmisji w następnym okresie TXOP. Potwierdzenia blokowe umożliwiają potwierdzenie całej transmisji zawartej w czasie TXOP w jednej ramce blokowego potwierdzenia BA (*Block Acknowledgement*). Gdy nadajnik chce odebrać ramkę ACK, wysyła ramkę żądania blokowego ACK i czeka na ramkę blokowego ACK. Po odebraniu potwierdzenia, nadajnik ponawia próbę przesłania uszkodzonych ramek albo w innym bloku, albo indywidualnie. Wyróżniamy dwa rodzaje ramek blokowego potwierdzenia – natychmiastowe oraz opóźnione. W pierwszym przypadku odbiorca otrzymujący żądanie potwierdzenia, wysyła go bezzwłocznie. Opóźnione Block ACK jest wysyłane w kolejnym uzyskanym czasie TXOP.

Standard IEEE 802.11ax wspiera trzy rodzaje potwierdzeń dla transmisji w dół oraz techniki OFDMA. Pierwszym jest potwierdzenie w formacie pojedynczego użytkownika. Obserwując Rys. 4 można zauważyć transmisję równoległą w dół, po której następuje wymiana ramek BAR/BA pomiędzy pojedynczą stacją, a punktem dostępowym.



Rys. 4. Przykład transmisji w dół z potwierdzeniami nadawanymi przez pojedynczych użytkowników [11]

Drugi rodzaj potwierdzenia jest uzyskiwany dzięki ramce wyzwalającej MU-BAR, która jest wysyłana jako pojedyncza ramka użytkownika (Rys. 5). Odpowiedzi BlockAck są wysyłane w ramkach TB PPDU.



Rys. 5. Przykład transmisji w dół z potwierdzeniami wyzwalanymi ramką MU-BAR [11]

W trzecim rodzaju potwierdzenia ramka MU-BAR jest zagregowana do ramki danych (Rys. 6). Odpowiedzi BlockAck są również wysyłane w ramkach TB PPDU.

Rys. 6. Przykład transmisji w dół z potwierdzeniami wyzwalanymi zagregowanymi ramkami MU-BAR [11]

W pracy porównana zostanie wydajność sieci korzystających z opisanych wyżej trybów potwierdzeń ramek.

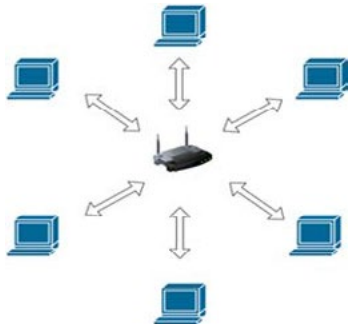
5. WYNIKI BADAŃ

Do symulacji scenariuszy badawczych wybrano narzędzie symulacyjne ns-3 w wersji 3.37 [10]. Narzędzie to pozwala na przeprowadzenie wysoko realistycznych symulacji sieci WLAN. Najważniejsze parametry symulacji wspólne dla wszystkich scenariuszy zebrano w Tabeli 1. Na wszystkich rysunkach błąd każdego punktu symulacji dla 95% przedziału ufności nie przekraczał ±2%.

Tab. 1. Wartości parametrów symulacji

Parametr	*Wartość*
Pasmo częstotliwości	5 [GHz]
Szerokość kanału	40 [MHz]
Poziom mocy nadawczej stacji	20 [dBm]
Poziom mocy nadawczej AP	20 [dBm]
Liczba anten	1
MCS	9 (256QAM 5/6)
Odległość AP od stacji	5 [m]
Okres ochronny	800 [ns]
RTS/CTS	Wył.
Fragmentacja	Wył.
Agregacja	A-MPDU
Protokół transportowy	UDP
Typ ruchu	CBR
Rozmiar pakietu	1000 B

W pierwszym scenariuszu symulacje zostaną przeprowadzone dla rosnącej liczby stacji od 2 do 50, dla pakietów o rozmiarze 1000 bajtów. Rys. 7 przedstawia analizowaną topologię sieciową. Drugi scenariusz dotyczy analizy metryk wydajnościowych sieci dla zwiększającej się wielkości pakietu zaczynając od 100 bajtów, a kończąc na 1400 bajtach. Dla tego przypadku zostanie przeanalizowana transmisja w łączu w dół dla 20 stacji.

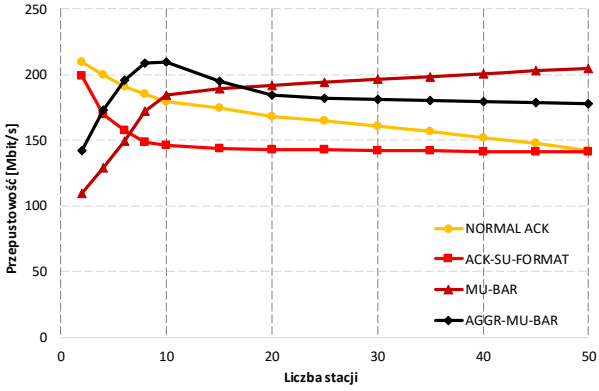


Rys. 7. Rozważana topologia sieciowa

5.1. Scenariusz 1 – zmienna liczba stacji

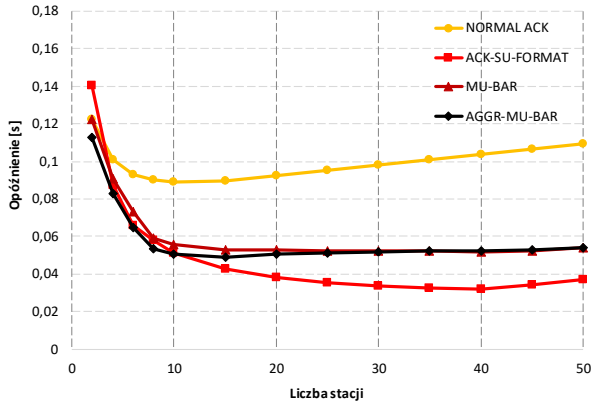
W tym scenariuszu został zbadany ruch DL OFDMA dla trzech typów potwierdzeń i transmisji DL OFDM ze zwykłym potwierdzeniem w funkcji rosnącej liczby stacji od 2 do 50. Całkowity ruch oferowany dla tego scenariusza wyniósł 230 Mbit/s. Został on podzielony po równo między stacje tak, aby sieć była w stanie nasycenia. Rys. 8 przedstawia całkowitą przepustowość sieci. Obserwując przebiegi w funkcji rosnącej liczby stacji najlepszą wydajność prezentuje potwierdzenie wyzwalane zagregowaną ramką MU-BAR, pomimo początkowej przewagi zwykłego ACK w transmisji OFDM. Drugim wydajnym potwierdzeniem jest potwierdzenie z ramką wyzwalającą MU-BAR. Słabszą wydajność przed-

stawia transmisja DL OFDMA z pojedynczym potwierdzeniem użytkownika, co jest spodziewanym wynikiem związanym ze sposobem implementacji mechanizmu potwierdzeń. Mechanizm ten marnuje zasoby radiowe przez wymuszanie szeregowej transmisji potwierdzeń od kolejnych stacji, przedzielanych odstępami międzyramkowymi SIFS. Dla niewielkiej liczby stacji transmisja równoległa staje się bardziej kosztowna przez istniejący w sieci nadmiar związany z nadawaniem większej liczby ramek kontrolnych. W przypadku transmisji OFDM wraz ze wzrostem liczby stacji przepustowość maleje jednostajnie. Zgodnie z oczekiwaniami, techniki potwierdzeń stosowane w MU OFDMA wykazują lepszą wydajność w środowiskach sieci gęstych.



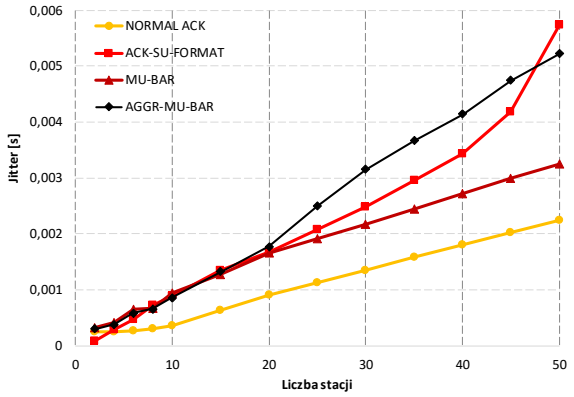
Rys. 8. Przepustowość sieci w funkcji liczby stacji dla różnych mechanizmów potwierdzeń

Mniejsze opóźnienia transmisji w przypadku stosowania potwierdzeń blokowych wynikają bezpośrednio z założeń techniki MU OFDMA. Punkt dostępowy przydziela użytkownikom jednostki zasobów, umożliwiając równoległe przesyłanie danych, co ma przełożenie na wielkość opóźnień w sieci (Rys. 9). Do pewnego momentu opóźnienie jest większe od średniej i wskazuje na nieefektywność trybu w mało zagęszczonych środowiskach. Powodem jest ponownie nieproporcjonalny do liczby stacji narzut sygnalizacyjny. Użytkownik w transmisji OFDM zajmuje cały dostęp do kanału w jednostce czasu, większe opóźnienia są spowodowane oczekiwaniem na zwolnienie kanału przez nadawcę. Im więcej stacji uczestniczy w transmisji tym większe jest opóźnienie.



Rys. 9. Opóźnienie transmisji w funkcji liczby stacji dla różnych mechanizmów potwierdzeń

Rys. 10 prezentuje jitter w funkcji rosnącej liczby stacji dla różnych mechanizmów opóźnień. Przeciwnie do samego opóźnienia, jego zmiany są generalnie większe dla transmisji DL OFDMA. Jednakże te kilku milisekundowe wartości nie wpływają znacząco na jakość transmisji, pomimo że potwierdzenia ze zagregowaną ramką MU-BAR oraz nadawane przez pojedynczych użytkowników wypadają w tym porównaniu najgorzej. Większe wartości jittera wynikają z konieczności synchronizacji w czasie transmisji równoległej do wielu stacji.

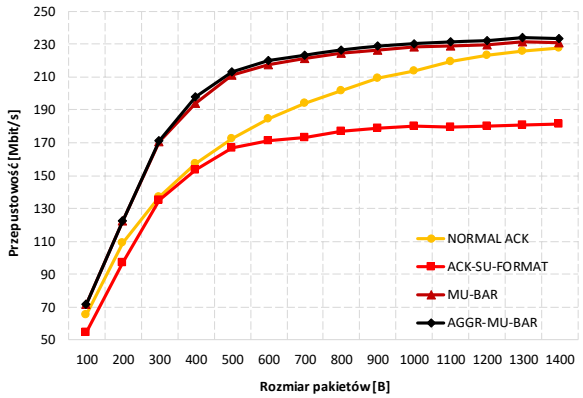


Rys. 10. Zmienność opóźnienia transmisji w funkcji liczby stacji dla różnych mechanizmów potwierdzeń

5.2. Scenariusz 2 – zmienny rozmiar pakietów

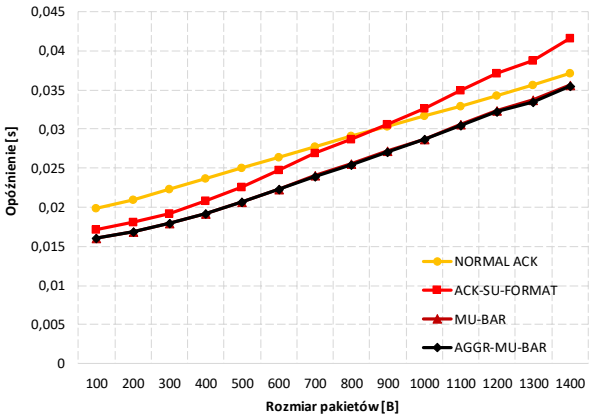
Drugi scenariusz porównuje efektywność mechanizmów potwierdzeń w funkcji zwiększającego się rozmiaru pakietów. AP nadawał ruch o wielkości 280 Mbit/s do 20 stacji. Został on rozdzielony po równo między stacje tak, aby sieć pracowała w stanie nasycenia.

Wyniki uzyskanej przepustowości dla średniej długości pakietów potwierdzają korzyść stosowania potwierdzeń blokowych w MU OFDMA (Rys. 11). Dopiero przy największych rozmiarach pakietów, transmisja OFDM jest w stanie konkurować z MU OFDMA. Nieefektywność pracy dla małych pakietów wynika z działania techniki OFDM, gdzie transmisja krótkich pakietów użytkownika związana jest z dużym nadmiarem jakim obarczona jest taka właśnie transmisja. Dla MU OFDMA dzięki dostępowi do kanału z użyciem RU, transmisja krótszych pakietów jest nieco bardziej wydajna. Najgorzej w zestawieniu wypadł mechanizm potwierdzeń przez pojedynczych użytkowników.



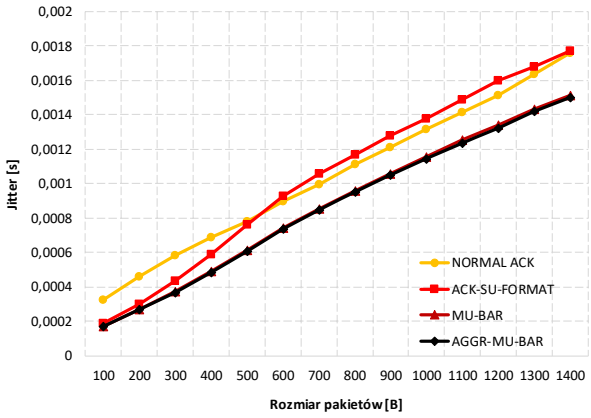
Rys. 11. Przepustowość sieci w funkcji rozmiaru pakietu dla różnych mechanizmów potwierdzeń

Transmisja długich pakietów zwiększa przepustowość, ale oznacza również większe opóźnienia (por. Rys. 12). Ponownie można zaobserwować nieco lepsze wyniki dla MU OFDMA, które również wynikają z liczby stacji uczestniczących w transmisji. Po raz kolejny łączenie pakietów wielu użytkowników i wysyłanie ich równolegle jest bardzie opłacalne. Przewaga MU OFDMA jest bardziej zauważalna dla krótkich pakietów, jednakże dla długich pakietów nieco się zmniejsza. W przypadku najdłuższych pakietów ponownie najgorzej radzi sobie mechanizm potwierdzeń nadawanych przez pojedynczych użytkowników. Związane jest to z szeregową transmisją kolejnych ramek potwierdzeń, blokujących przez dłuższy czas dostęp do kanału radiowego i uniemożliwiających realizację kolejnej transmisji.



Rys. 12. Opóźnienie transmisji w funkcji rozmiaru pakietu dla różnych mechanizmów potwierdzeń

Wariancja opóźnienia pokazana na Rys. 13 również ujawnia podobne trendy co wykresy opóźnienia. Niezależnie od rozmiaru pakietu, najmniejszy jitter osiąga sieć dla potwierdzeń wieloużytkownikowych MU OFDMA z wyłączeniem mechanizmu potwierdzeń nadawanych przez pojedynczych użytkowników.



Rys. 13. Zmienność opóźnienia transmisji w funkcji rozmiaru pakietu dla różnych mechanizmów potwierdzeń

6. PODSUMOWANIE

Celem pracy była analiza wydajności działania mechanizmów potwierdzeń ramek dla trybu MU OFDMA sieci standardu IEEE 802.11ax. Analiza zebranych wyników dowiodła, że tryby potwierdzeń wieloużytkownikowych są szczególnie polecane w sieciach złożonych ze średniej i dużej liczby stacji. W przypadku kiedy w sieci znajduje się ok. 20 stacji, największy zysk jest osiągany dla pakietów o wielkości ok. 300-800 bajtów. Technika OFDM z tradycyjną metodą potwierdzeń jest nadal efektywna dla małej liczby stacji lub gdy w sieci przesyłane są bardzo długie pakiety. Dalsza analiza badanego tematu powinna objąć możliwość transmisji z priorytetyzacją, czyli z podziałem na klasy ruchu EDCA: Voice, Video, Best Effort, Background.

PODZIĘKOWANIA

Praca została sfinansowana przez Ministerstwo Nauki i Szkolnictwa Wyższego w ramach subwencji dla Wydziału Informatyki, Elektroniki i Telekomunikacji AGH.

LITERATURA

[1] IEEE Standard for Information Technology--Telecommunications and Information Exchange between Systems - Local and Metropolitan Area Networks--Specific Requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications - Redline," in IEEE Std 802.11-2020 (Revision of IEEE Std 802.11-2016) - Redline, vol., no., pp.1-7524, 26 Feb. 2021.

[2] G. Naik, S. Bhattarai and J. -M. Park, *Performance Analysis of Uplink Multi-User OFDMA in IEEE 802.11ax*, 2018 IEEE International Conference on Communications (ICC), 2018, pp. 1-6.

[3] Lee, K.-h. *Performance Analysis of the IEEE 802.11ax MAC Protocol for Heterogeneous Wi-Fi Networks in Non-Saturated Conditions*, Sensors 2019, 19, 1540.

[4] M. Ş. Kuran, A. Dilmac, Ö. Topal, B. Yamansavascilar, S. Avallone and T. Tugcu, *Throughput-maximizing OFDMA Scheduler for IEEE 802.11ax Networks*, 2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications, 2020, pp. 1-7.

[5] K. -H. Lee, *Using OFDMA for MU-MIMO User Selection in 802.11ax-Based Wi-Fi Networks*, in IEEE Access, vol. 7, pp. 186041-186055, 2019.

[6] R. Faruqui and S. Ghani, *A simulation study of Block Acknowledgements and TXOPs under varying channel conditions*, 2008 IEEE International Multitopic Conference, 2008, pp. 286-289.

[7] Tinnirello and Sunghyun Choi, *Efficiency analysis of burst transmissions with block ACK in contention-based 802.11e WLANs*, IEEE International Conference on Communications, 2005. ICC 2005. 2005, 2005, pp. 3455-3460.

[8] Cisco Meraki - Wi-Fi 6 (802.11ax) Technical Guide https://documentation.meraki.com/MR/Wi-Fi_Basics_and_Best_Practices/Wi-Fi_6_(802.11ax)_Technical_Guide

[9] Ns-3 Design Documentation – Model Library: https://www.nsnam.org/docs/models/html/wifi-design.html

[10] Ns-3 Home Page: https://www.nsnam.org/

WDROŻENIE KONTROLERA SDN DO ZARZĄDZANIA PUNKTAMI DOSTĘPOWYMI STANDARDU IEEE 802.11

SDN CONTROLLER DEPLOYMENT FOR IEEE 802.11 ACCESS POINT MANAGEMENT

Anita Jamroży[1]; Katarzyna Kosek-Szott[2];

[1] Akademia Górniczo-Hutnicza im. Stanisława Staszica, Instytut Telekomunikacji, Kraków, ajamrozy@o2.pl
[2] Akademia Górniczo-Hutnicza im. Stanisława Staszica, Instytut Telekomunikacji, Kraków, kks@agh.edu.pl

DOI: 10.15199/59.2023.4.38

Streszczenie: **Artykuł przedstawia wstępne wyniki wdrożenia kontrolera SDN do zarządzania punktami dostępowymi sieci standardu IEEE 802.11. Skorzystano z architektury 5g-EmPower oraz przeprowadzono badania przepustowości sieci działającej w oparciu o to rozwiązanie. Wyniki badań pokazały, że wydajność takiej sieci jest niska oraz, że zarządzanie sieciami Wi-Fi oparte na kontrolerach SDN wymaga dalszych badań. Dodatkowo, w artykule przedstawiono przegląd dostępnych rozwiązań SDN przeznaczonych dla sieci Wi-Fi.**

Abstract: **The article presents the results of the deployment of an SDN controller for managing IEEE 802.11 network access points. The 5g-EmPower architecture was used and network capacity was tested for this solution. The conducted research showed that the performance of the network based on the 5g-EmPower SDN controller is low, and managing WiFi networks with SDN controllers should be further investigated. In addition, the article presents an overview of available SDN controller solutions for IEEE 802.11 wireless networks.**

Słowa kluczowe: **5g-EmPower, IEEE 802.11, kontroler, punkt dostępowy, Software Defined Networking**

Keywords: **5g-EmPower, access point, controller, IEEE 802.11, Software Defined Networking**

1. WSTĘP

Sieci bezprzewodowe standardu IEEE 802.11 stanowią ważny element w kontekście dostępu do Internetu. Liczba urządzeń korzystających z łączy bezprzewodowych nieustannie rośnie. Są to m.in. telefony, laptopy, czy urządzenia Internetu Rzeczy. Administratorzy sieci bezprzewodowych są świadomi rosnącego obciążenia i starają się wprowadzać rozwiązania pozwalające na zwiększenie niezawodności i wydajności zarządzanych sieci. Aby zapewnić dobrą jakość sygnału na dużych powierzchniach (a tym samym większe przepustowości), administratorzy rozmieszczają wiele punktów dostępowych, które obsługują tą samą sieć. Powoduje to potrzebę implementacji rozwiązań, które pozwalają na efektywne korzystanie z tej sieci. Są to m.in. algorytmy doboru kanałów i mocy, mechanizmy przełączania użytkowników (ang. *user handoffs*), czy priorytetyzacja ruchu w oparciu o metryki QoS (ang. *Quality of Service*).

SDN (ang. *Software Defined Networking*) jest podejściem, które może być dodatkowo pomocne przy wdrażaniu rozległych sieci bezprzewodowych. Bazuje ono na centralizacji zarządzania siecią w postaci jednostki zwanej kontrolerem. Pozwala to m.in. na precyzyjny dobór trasy dla przepływów sieciowych oraz kontrolowany przydział zasobów. SDN wychodzi także naprzeciw wdrażaniu mechanizmów wspomnianych wcześniej. Dodatkowo ułatwia proces zarządzania sieciami teleinformatycznymi.

Bezprzewodowy SDN (ang. *wireless* SDN, WSDN) nie jest jeszcze tak popularnym rozwiązaniem, jak SDN dla sieci przewodowych. W literaturze istnieją jednak prototypy rozwiązań, które umożliwiają wdrożenie kontrolera SDN np. dla sieci Wi-Fi. Często są to niestety rozwiązania niestabilne. Niniejszy artykuł prezentuje przegląd dostępnych rozwiązań WSDN oraz wstępne wyniki testów przeprowadzonych dla środowiska działającego w oparciu o architekturę 5g-EmPower [7].

Artykuł składa się z następujących części. Sekcja 2 wyjaśnia koncepcję WSDN. Sekcja 3 prezentuje przegląd rozwiązań WSDN dedykowanych dla sieci IEEE 802.11. Sekcja 4 przedstawia proces wdrożenia kontrolera SDN 5g-EmPower. Sekcja 5 przedstawia wyniki przeprowadzonych testów. W sekcji 6 zawarto najważniejsze wnioski z przeprowadzonych badań.

2. BEZPRZEWODOWY SDN

SDN jest paradygmatem umożliwiającym scentralizowane zarządzanie sieciami teleinformatycznymi oraz dopasowanie ich działania do potrzeb użytkowników. WSDN skupia się przede wszystkim na elementach charakterystycznych dla sieci bezprzewodowych, które często nie są implementowane w klasycznych rozwiązaniach SDN. Systemy bezprzewodowe posiadają zmienną charakterystykę łącza radiowego, mobilne terminale oraz zmienną topologię. Kontroler sieci bezprzewodowej powinien więc dostosowywać się do tych zmian. Dodatkowo, operatorzy sieci bezprzewodowych mogą chcieć podzielić dostępną sieć na logiczne segmenty, które pozwoliłyby sprecyzować różnice w sposobie obsługi wspieranych usług. W przypadku sieci IEEE 802.11 operacja ta została określona terminem *Wi-Fi slicing* [3]. *Wi-Fi slicing* pozwala na efektywne podzielenie całej sieci oraz monitorowanie jej przez centralny kontroler. *Wi-Fi*

Slicing został zaimplementowany w kontrolerze 5g-EmPower [7]. Jest to możliwe dzięki implementacji wirtualnego punktu dostępowego o indywidualnym BSSID (ang. *Basic Service Set Identifier*) dla każdego urządzenia mobilnego. Dodatkowo pakiety sieciowe są dzielone na podstawie wartości DSCP (ang. *differentiated services code point*) oraz przydzielane do wcześniej zdefiniowanych kolejek. Administrator sieci ustawia na kontrolerze maksymalną długość czasu trwania transmisji dla każdej kolejki. Pozwala to na precyzyjne zarządzanie obsługiwanym ruchem teleinformatycznym.

3. PRZEGLĄD ROZWIĄZAŃ WSDN

Ponieważ SDN jest obecnie bardzo popularnym kierunkiem badań, w literaturze znajdują się rozwiązania realizacji kontrolera SDN dla sieci Wi-Fi. Są to m. in. MonFi [4], SWIFT [11], Ethanol [1], oraz 5g-EmPower [7].

MonFi jest programem pozwalającym na dokładne monitorowanie sieci WiFi. Nie jest to klasyczne rozwiązanie jeśli chodzi o kontrolery SDN, jednakże stanowi dobrą podstawę do wdrożenia tej architektury. Według autorów MonFi bazuje na zmodyfikowanych sterownikach *ath9k* oraz *ath10k,* które umożliwiają zebranie wszystkich niezbędnych informacji, nawet niskopoziomowo.

SWIFT jest projektem zapewniającym zarządzanie przepływami za pomocą kontrolera SDN. Jego program oparty jest o model Ryu [8]. SWIFT jest przeznaczony do zarządzania różnymi typami punktów dostępowych (od urządzeń przemysłowych takich jak Cisco, po urządzenia domowe z zainstalowanym otwartym systemem OpenWRT) [11]. Jest to ciekawe rozwiązanie, ponieważ zróżnicowany sprzęt może przysparzać wielu problemów w procesie administracji.

Kolejnym ważnym projektem jest projekt Ethanol, który zapewnia dedykowane programy dla kontrolera SDN oraz dla punktu dostępowego. Dodatkowo autorzy zapewnili wiele funkcji, które pozwalają na zwiększenie wydajności oraz bezpieczeństwa sieci. Są to m.in. lokalizacja i przełączanie użytkowników pomiędzy punktami dostępowymi wraz z wcześniejszym uwierzytelnieniem, oraz implementacja zasad QoS w oparciu o rozszerzenie IEEE 802.11e.

Ostatnim ważnym rozwiązaniem jest architektura 5g-EmPower. Jest to bardzo złożony projekt oferujący zarówno kontroler dla sieci Wi-Fi, jak i zarządzanie stacjami dla sieci mobilnych 5G. Kontroler bazuje na środowisku napisanym w języku Python oraz serwerze Tornado Web Server [13]. W ramach projektu stworzono również specjalne oprogramowanie, które instalowane jest na punkcie dostępowym. Bazuje ono na zmodyfikowanej wersji systemu operacyjnego OpenWRT, która została wzbogacona o oprogramowanie Click [12] i Open VSwitch [5]. Oba programy umożliwiają komunikację z kontrolerem SDN oraz realizację jego instrukcji.

4. WDROŻENIE KONTROLERA WSDN

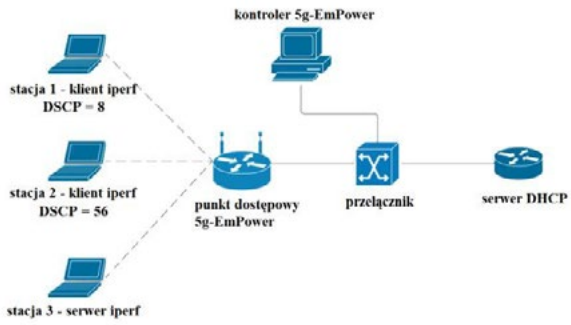
Z dostępnych rozwiązań WSDN przedstawionych w sekcji 3, jedynie architektura 5g-EmPower uruchamia się poprawnie. Podczas wdrażania rozwiązań MonFi, SWIFT, oraz Ethanol natrafiono na liczne błędy w oprogramowaniu, które uniemożliwiły ich dalsze testowanie. Zaobserwowane problemy wynikały głównie z przestarzałych wersji wykorzystywanych języków programowania (np. języka Python) oraz niewspieranych już dystrybucji systemu operacyjnego Linux.

Kontroler 5g-EmPower wymaga użycia jednego komputera z zainstalowanym oprogramowaniem kontrolera, co najmniej jednego punktu dostępowego, oraz stacji klienckich. Zgodnie z dokumentacją projektu [7] do instalacji kontrolera WSDN wymagany jest komputer z najnowszą dystrybucją systemu operacyjnego Linux. Do instalacji oprogramowania punktu dostępowego potrzebny jest ruter z układem scalonym Atheros.

W zestawionym środowisku testowym Jako kontrolera użyto komputer stacjonarny z systemem operacyjnym Linux Ubuntu 20.04.5 LTS, a jako punkt dostępowy wykorzystano urządzenie TP-Link WDR4300. Dodatkowo do przeprowadzenia pomiarów użyto trzech stacji klienckich - laptop Dell Latitude E7440, laptop HP Zbook 14 G8 i laptop HP Zbook 15 G3. Wszystkie maszyny miały zainstalowane oprogramowanie Windows 10. Do stacji klienckich podłączono karty bezprzewodowe sieciowe Edimax AC1200. Z racji różnic sprzętowych w stacjach klienckich zdecydowano się na użycie tego samego modelu kart w każdym urządzeniu

Instalacja oprogramowania 5g-EmPower odbyła się w oparciu o instrukcję [9, 10]. Podczas instalacji skorzystano z gotowego obrazu systemu operacyjnego dla punktu dostępowego TP-Link WDR4300 [6].

Po zakończeniu procesu instalacji, zbudowano sieć o topologii przedstawionej na rysunku 1. Zestawiono sieć w oparciu o rozszerzenie IEEE 802.11n. Stacje przesyłały ruch używając MCS (*modulation and coding scheme*) 4. Do pomiarów przepustowości sieci użyto programu iperf v3.1.3 [2].



Rys. 1. Topologia środowiska testowego 5g-EmPower

Wydajność zestawionej sieci została przeanalizowana dla czterech scenariuszy. Pomiary obejmowały ruch TCP i UDP oraz różne wartości parametru *Wi-Fi Slice* (por. sekcja 2). *WiFi Slice* używało kolejkowania zgodnie z algorytmem *Weighted round robin.* Należy zaznaczyć, że wartości WiFi Slice obowiązują jedynie dla ruchu wychodzącego z punktu dostępowego. Nie zastosowano agregacji pakietów. Ruch

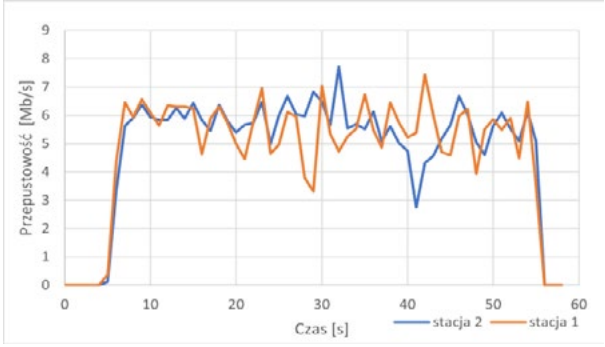
przesyłany od klienta iperf do serwera. Wielkość pakietów wynosiła odpowiednio 120 KB dla TCP i 1.5 KB dla UDP. Za pomocą programu iperf ustawiono dwie wartości kodów DSCP – 8 i 56. W tabeli 1 przedstawiono scenariusze testowe oraz wartości parametru *Wi-Fi Slice* przypisane do każdego z badanych przepływów.

Tab. 1. Wartości Wi-Fi Slice powiązane z różnymi wartościami DSCP (8 i 56) dla protokołu TCP i UDP.

Scenariusz	*Rodzaj ruchu*	*Wi-Fi Slice [µs]*	
		DSCP = 8	*DSCP = 56*
1	TCP	12 000	12 000
2	TCP	12 000	5000
3	UDP	12 000	12 000
4	UDP	12 000	5000

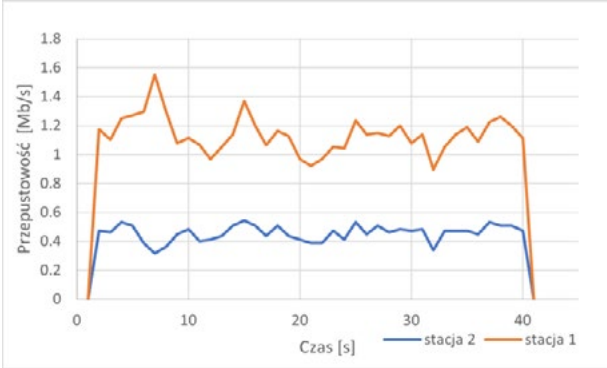
5. WYNIKI

W pierwszym scenariuszu parametr *Wi-Fi Slice* miał taką samą wartość dla obu przepływów (12000 *µs*), czyli obu klasom ruchu (DSCP = 8 i DSCP = 56) przypisano taką samą ilość dostępnych zasobów radiowych. Uzyskane statystyki przepustowości przedstawiono na rysunku 2. Oba przepływy osiągnęły podobne wartości przepustowości. Jest to zgodne z ustawionymi wartościami parametru *Wi-Fi*, co potwierdza poprawne wdrożenie kontrolera. Średnia prędkość transmisji dla obu stacji wynosi około 5,.2 Mb/s.



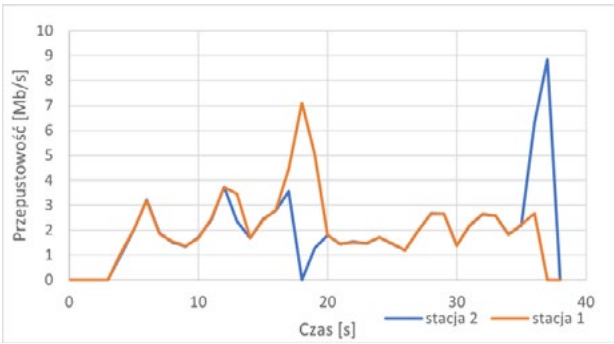
Rys. 2. Pomiary przepustowości dla scenariusza 1: Wi-Fi Slice = 12000 µs dla DSCP = 8 i DSCP = 56, ruch TCP

W scenariuszu 2, parametr *Wi-Fi Slice* dla ruchu z DSCP = 8 został skonfigurowany na ponad dwukrotnie większą wartość niż dla ruchu z DSCP = 56. Wyniósł on odpowiednio 12000 µs i 5000 µs. Statystyki przepustowości przedstawiono na rysunku 3. W tym przypadku przepustowości obu przepływów znacząco różnią się od siebie. Stacja 2 transmitująca ruch z DSCP = 56 osiągnęła przepustowość wynoszącą około 0,4 Mb/s. Stacja 1, transmitująca w tym czasie z wyższym priorytetem, osiągnęła przepustowość o wartości średniej 1,2 Mb/s. Różnica w prędkości transmisji pomiędzy obiema stacjami jest około trzykrotna. Jest to wynik, który w dużej mierze pokrywa się z ustawionymi wartościami parametru *Wi-Fi Slice* na kontrolerze.



Rys. 3. Pomiary przepustowości dla scenariusza 2: Wi-Fi Slice = 12000 µs dla DSCP = 8 oraz Wi-Fi Slice = 5000 µs dla DSCP = 56, ruch TCP

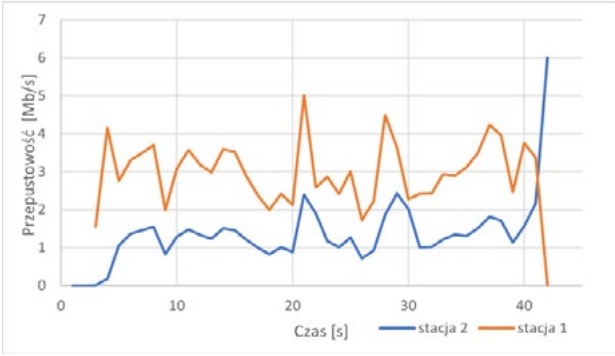
Dla protokołu UDP, w scenariuszu 3, wartości parametru *Wi-Fi Slice* miały tę samą wartością (12000 *µs*) dla obu przepływów (z DSCP = 8 i DSCP = 56). W programie iperf ustawiono dla każdej ze stacji prędkość nadawania na 10 Mb/s aby uzyskać pełne nasycenie sieci. Zaobserwowano częściowe straty pakietów. Wyniki dla tego scenariusza przedstawiono na rysunku 4.



Rys. 4. Pomiary przepustowości dla scenariusza 3: Wi-Fi Slice = 12000 µs dla DSCP = 8 i DSCP = 56, ruch UDP

W scenariuszu 3 przepustowość dla obu stacji jest podobna i wynosi około 2.2 Mb/s. Pojawiają się jednak przerwy w transmisji. Dodatkowo, obserwowana wartość przepustowości jest znacząco niższa niż w przypadku ruchu TCP (w scenariuszu 1). Jest to wynik niezgodny z intuicją, który wymaga dalszych badań.

W scenariuszu 4, parametr *Wi-Fi Slice* dla DSCP = 8 został ustawiony na wartość 12000 µs, a dla DSCP = 56 na wartość 5000 µs. Statystyki przepustowości dla tego scenariusza przedstawiono na rysunku 5. W tym przypadku średnie wartości przepustowości są różne od siebie dla obu stacji i wynoszą odpowiednio ok. 2,7 Mb/s dla stacji 1 i 1,2 Mb/s dla stacji 2. Biorąc pod uwagę ustawienia parametru *Wi-Fi Slice* są to wartości w dużej mierze pokrywające się z ustawieniami wprowadzonymi na kontrolerze. W porównaniu do scenariusza 2, wartości otrzymywane dla ruchu UDP są wyższe niż dla ruchu TCP. Odpowiada to charakterystyce obu protokołów.



Rys. 5. Pomiary przepustowości dla scenariusza 4: Wi-Fi Slice skonfigurowane na 12000 µs dla DSCP = 8 i 5000 µs dla DSCP = 56, ruch UDP.

We wszystkich badanych przypadkach, uzyskiwane wartości przepustowości są dużo niższe od spodziewanych. Ograniczenie to może wynikać z dużego narzutu komunikacyjnego pomiędzy kontrolerem a punktem dostępowym lub nieprawidłowej implementacji protokołu sygnalizacyjnego. Ta hipoteza wymaga dalszych badań.

6. WNIOSKI

Celem przeprowadzonych prac było wdrożenie kontrolera SDN do zarządzania siecią standardu IEEE 802.11. Cel udało się osiągnąć, używając oprogramowania 5g-EmPower. Testowano również cztery inne rozwiązania - MonFi, SWIFT i Ethanol. Niestety ich wdrożenie zakończyło się niepowodzeniem.

Do analizy działania architektury 5g-EmPower wykorzystano parametr *Wi-Fi Slice*, który stanowi integralną część projektu 5g-EmPower [7]. Przetestowano różne konfiguracje parametru *Wi-Fi Slice* dla dwóch protokołów transportowych (TCP i UDP).

W badanych scenariuszach, wartość parametru *Wi-Fi Slice* wpływała na rozkład przepustowości pomiędzy przepływami posiadającymi różne wartości kodów DSCP. Obserwowana różnica była zauważalna i zgodna z oczekiwaniami. Wystąpiły jednak niespodziewane różnice w działaniu kontrolera dla ruchu TCP i UDP (por. scenariusze 1 i 3). Dodatkowo, zaobserwowano, że osiągane wartości przepustowości są mocno ograniczone. Dlatego wdrożona architektura wymaga dalszych badań. Powinny one się opierać na inwestygacji problemu z niską przepustowością oraz w miarę możliwości rozwiązania problemu.

Podsumowując, praca dowodzi, że rozwiązania oparte na kontrolerze SDN dla sieci Wi-Fi istnieją i mogą być wdrażane. Obecne rozwiązania są jednak często niedoskonałe. Dodatkowo, większość z istniejących kontrolerów WSDN jest zależna od stosowanych platform sprzętowych, wersji systemów operacyjnych i języków programowania, co utrudnia ich weryfikację oraz przyszłe wdrożenie w rzeczywistych sieciach.

LITERATURA

[1] Moura, Henrique, Gabriel VC Bessa, Marcos AM Vieira, and Daniel F. Macedo. 2015. "Ethanol: Software defined networking for 802.11 wireless networks." In 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM): pp. 388-396.

[2] Gueant, Vivien. “IPerf - The TCP, UDP and SCTP Network Bandwidth Measurement Tool,” n.d. https://iperf.fr/.

[3] Coronado, Estefanía, Roberto Riggio, José Villa1ón, and Antonio Garrido. 2018. "Lasagna: Programming abstractions for end-to-end slicing in software-defined wlans." In 2018 IEEE 19th International Symposium on" A World of Wireless, Mobile and Multimedia Networks"(WoWMoM): pp. 14-15.

[4] Sheth, Jaykumar, and Behnam Dezfouli. 2021. "MonFi: A Tool for High-Rate, Efficient, and Programmable Monitoring of WiFi Devices." In 2021 IEEE Wireless Communications and Networking Conference (WCNC): pp. 1-7.

[5] “Open VSwitch.” n.d. https://www.openvswitch.org/.

[6] Riggio, Roberto. “Pre Built WTP Firmwares.” GitHub, n.d. https://github.com/5g-empower/5g-empower.github.io/wiki/Pre-built-WTP-Firmwares.

[7] Riggio, Roberto, Mahesh K. Marina, Julius Schulz-Zander, Slawomir Kuklinski, and Tinku Rasheed. 2015 "Programming abstractions for software-defined wireless networks." IEEE Transactions on Network and Service Management 12, no. 2 (2015): 146-162.

[8] “Ryu SDN Framework,” n.d. https://ryu-sdn.org/.

[9] Riggio, Roberto. “Setting up the Controller.” GitHub, n.d. https://github.com/5g-empower/5g-empower.github.io/wiki/Setting-up-the-Controller.

[10] Riggio, Roberto. “Setting up the WTP.” GitHub, n.d. https://github.com/5g-empower/5g-empower.github.io/wiki/Setting-up-the-WTP.

[11] Hätönen, Seppo, Petri Savolainen, Ashwin Rao, Hannu Flinck, and Sasu Tarkoma. 2018. "SWIFT: bringing SDN-Based flow management to commodity Wi-Fi access points." In 2018 IFIP Networking Conference (IFIP Networking) and Workshops, pp. 1-9

[12] Kohler, Eddie, Robert Morris, Benjie Chen, John Jannotti, and M. Frans Kaashoek. 2000. "The Click modular router." ACM Transactions on Computer Systems (TOCS) 18, no. 3 (2000): 263-297.

[13] “Tornado Web Server — Tornado 6.3.1 Documentation,” n.d. https://www.tornadoweb.org/en/stable/.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

ATAKI PRZECIWKO KLASYFIKACJI ZADAŃ TRANSMISYJNYCH I OBLICZENIOWYCH ORAZ METODY ICH WYKRYWANIA

ATTACKS AGAINST COMMUNICATION AND COMPUTING TASK CLASSIFICATION AND MITIGATION TECHNIQUES

Younes Salmi[1]; Hanna Bogucka[1]

[1] Instytut Radiokomunikacji, Politechnika Poznańska, Poznań,
younes.salmi@student.put.poznan.pl; hanna.bogucka@put.poznan.pl

DOI: 10.15199/59.2023.4.39

Streszczenie: **Algorytmy klasyfikacji oparte na uczeniu maszynowym umożliwiają zoptymalizowaną alokację zadań telekomunikacyjnych i obliczeniowych (2C) do serwerów brzegowych sieci. W artykule omówiono ataki zatruwające, które mają negatywny wpływ na klasyfikację zadań 2C w obszarach, w których granica decyzyjna jest niejasna. Proponujemy metodę trenowania modelu oraz wykorzystanie testu Silhouette do wykrywania i unikania ataków. Wykazujemy skuteczność tych metod wobec rozważanych ataków.**

Abstract: **Machine learning-based classification algorithms allow communication and computing (2C) task allocation to network edge servers. This article considers poisoning of classifiable 2C data features in two scenarios: noise-like jamming and targeted data falsification. These attacks have a fatal effect on classification in the feature areas with unclear decision boundary. We propose training and noise detection using the Silhouette Score to detect and mitigate attacks. We demonstrate effectiveness of our methods.**

Słowa kluczowe: **klasyfikacja danych, ataki zatruwające, przetwarzanie brzegowe, uczenie maszynowe.**

Keywords**: clustering, poisoning attack, edge computing, machine learning.**

1. INTRODUCTION

Edge intelligence refers to intelligent services and functionalities implemented at the network's edge, which require novel architectures and communication-efficient training methods. These designs introduce new issues such as limited access to training data, low inference accuracy, lack of generalization, and processing power and memory limitations for edge devices. Below, we address edge computing network architecture with diverse servers and IoT devices.

Optimizing communication and computing (2C) task allocation to the appropriate servers is challenging. In this work, we investigate employing Machine Learning (ML) to classify generated 2C requests. This classification is intended to assist near-optimal request delegation to an appropriate server at the network's edge.

The shared nature of the wireless transmission medium makes it prone to attacks; in particular, the attacker may target the dataset, the prediction, or the ML model itself. The attacks on traffic steering may lead to fatal errors for sensitive applications, e.g., those requiring ultra-low latency and ultra-high reliability (URLLC). Here, we consider 2C task classification based on the following required QoS features: end-to-end latency, the reliability described by the packet error rate PER, and the computational complexity of the delegated task. The classification criteria may change, but in the case of multi-class classification with defined sensitivity levels (from non-critical to highly critical tasks), any misclassification may result in lost calls for edge computing tasks.

Below, we investigate how data poisoning attacks impact 2C task classification and offloading. In Section 2, we describe ML-based classification algorithms and consider security threats to these algorithms. In Section 3, we investigate the simulation scenarios and discuss the achieved results. In Section 4, we conclude the work.

2. ML FOR AUTONOMOUS 2C TASKS CLASSIFICATION AND THREATS

Our work focuses on self-learning networks and systems that can autonomously manage resources and control functions. The usage patterns of the different 5G/6G services are highly stochastic; as a result, unsupervised learning clustering might be used to discover the patterns. In our consideration, data representing the 2C tasks have features reflecting the required Quality of Service (QoS). Moreover, 2C task clustering is based on their proximity in the feature space,=. The 5G/6G constraints limit the choice of the algorithm. For large datasets, clustering algorithms such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [1], hierarchical clustering, etc., are computationally expensive, hence may fail to satisfy the ultra-low-latency and the computational efficiency requirements. To address this issue, *k*-means is a fast and efficient algorithm for large datasets [2]. However, implementing *k*-means on 2C tasks with non-spherical clusters is impractical. In addition, the algorithm requires the number of clusters to be specified and is sensitive to the choice of initial clusters. Thus, additional tools for its application must be considered.

After the clusters are obtained, new requests using supervised learning are predicted. The technique used here must be adaptive to changes made by *k*-means after each single epoch. Moreover, the lack of a training dataset limits the considerations for only non-training phase techniques. Here, the choice is on *k'*-nearest neighbours (*k'*-NN) [3], where the new requests are assigned to the nearest cluster based on a metric such as Euclidean distance.

The algorithm assigns each example based on the majority voting of these neighbours, but the technique is sensitive to the number of *k'* neighbours. If a lower *k'* is considered, the *k'*-NN may overfit the training data. However, higher *k'* results in higher computational complexity.

Management of the 2C task delegation is vulnerable to attacks due to the shared nature of wireless medium. Attackers can use adversarial machine learning (AML) techniques to generate poisoning examples [4] and inject them into the system or use an RF noise source to alternate the dataset. This can lead to wrong allocation, overloading some servers, and inability to serve users.

3. STUDY SCENARIO

We consider edge computing network architecture with diverse servers (with diverse computing and storage capabilities) and diverse IoT devices (generating diverse communication and computing requests). This architecture is presented in Fig. 1.

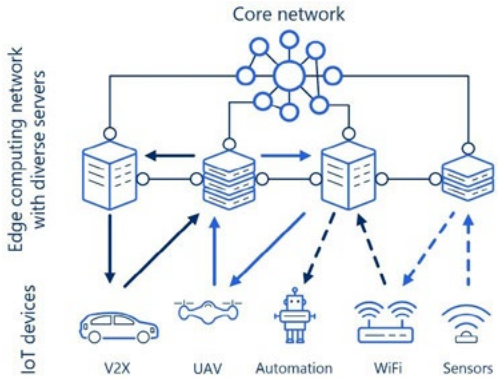


Fig. 1. System architecture

Out of multiple scenarios of requested 2C tasks, we have selected some representative use cases regarding expected latencies and the reliabilities (in terms of PER) the values of which are selected according to the QoS of 5G defined by ETSI in "(2022-05)1683GPP TS 23.501 version 17.4.0 Release 17". Moreover, the complexity of tasks is generated in the range of [200 Giga, 5000 Giga] in terms of floating operation points (FLOPs).

3.1. Dataset generation and scaling

The dataset is of three features representing the latency, the reliability, and the computational complexity respectively, and it consists of 10000 examples obtained after the first epoch, the events follow the Poisson process for the arrival time. To better distribute the clusters we consider features scaling, where the latency and computational complexity examples x^i are mapped using a function as in (1), while the reliability feature is logarithmically scaled before feeding the results to (1).

$$f(x^i) = c + \left(\frac{d-c}{b-a}\right)(x^i - a) \qquad (1)$$

The function returns the scaled example $f(x^i)$, where a and b are the pre-scaled features' minimum and maximum values respectively, while c and d are the post-scaled features' minimum and maximum values respectively (here we choose $c = 0$ and $d = 100$). The Silhouette score[1] on the examples scaled using (1) is the highest for $k = 8$.

3.2. Clustering and prediction

Given the unlabelled dataset, we chose k-means to run for its clustering. The u_k centroids are initialized randomly. The cost function is the sum of the squared distances of each data point to its assigned cluster. The decision of conversion to a new cluster is taken for the mean squared error $\varepsilon \leq 10^{-15}$ between the successive iterations' centroids. The results depend on the initial conditions. For different centroid feature values, different clustering might be obtained. Averaging over 10000 runs solves this problem. The final clustering results are shown in Fig. 2.

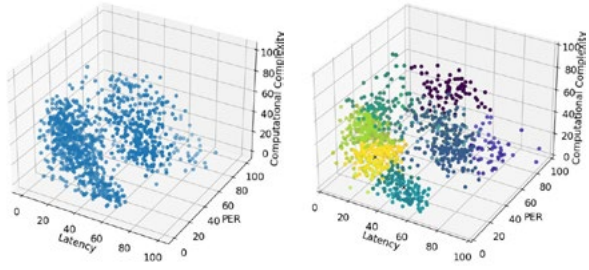


Fig.2. 3D visualization, (a) the post-processed dataset, (b) the k-means clustering results

Once the first epoch elapses and the examples are clustered, the final centroids are used as labels for prediction with *k'*-NN. The distance metric is chosen to be the Euclidean distance. To prevent overfitting, the training dataset is divided into training and cross-validation (CV) subsets, with proportions of 80% and 20% respectively. In addition, the CV dataset is used to optimize the neighbours number $k' = 50$ maximizing the F1-score [6] (The maximum F1 score = 0.97)[2].

To test the efficiency of ML-based task classification, it has been compared to round-robin task allocation. The number of clusters $k = 8$ has been the same as the number of servers. Each server has the 224i GFlops/s power (where i is a server's ID), and the same 2C latency and reliability. The scheduling of tasks in CPU is done as in [7]. The results have shown that the ML-based solution outperforms the round-robin it terms of the number of lost calls and the tasks processing time.

4. POISONING ATTACKS

In our scenario, we account for poisoning attacks which are considered as data injection with falsified examples to modify the dataset. The purpose of the attacks is to generate a bias for 2C task classification and allocation to the edge-computing servers, what may cause their overloading and drop the assigned tasks (or calls).

We consider two attacks. The first one (called *attack 1*) uses a malicious source that spoofs legitimate identity.

[1] The Silhouette Score is a measure of how similar an object is to its own cluster compared to other clusters [6].

[2] The F1-score combines the precision and recall of a classifier and is used to compare the classifiers [7].

The examples this source injects are grouped to form a cluster, positioned randomly in the dataset plane. The effects of this attack are observed on *k*-means and *k'*-NN algorithms (where 10000 tasks are predicted), and the system performance. For the second type of attack (called *attack 2*), the poisoning examples are not correlated, but they are perturbed versions of the legitimate examples. The examples are injected between the clusters with adjacent decision regions.

To evaluate the attacks' effect on the system performance different 2C links between end-users and servers (or VMs) are defined. Each link has three 2C metrics: latency, reliability, and computational complexity. For testing, the value of each metric is obtained by the *k*-means results. The task is said to be unacceptable if latency and/or the computational complexity is higher than the associated link value, and/or reliability is lower than the associated link value. The system performance is tested based on the average number of unacceptable tasks. The tasks to be classified are generated near the classifiers separating regions. In our simulations, we assumed a simplified scenario of assigning the tasks to offered services (2C links) by assigning a binary value of high or low to each of the requested (based on clustering) or offered features: latency, reliability, and computational complexity.

The results of data poisoning presented in Fig. 3 show that the malicious examples significantly disrupt the *k*-means and *k'*-NN algorithms and decrease the system performance for the higher number of injected examples. The worst results in terms of prediction accuracy are obtained for lower *k'* values. Moreover, attack 2 results in a lower clustering and prediction accuracy but also a lower number of dropped calls compared to attack 1.

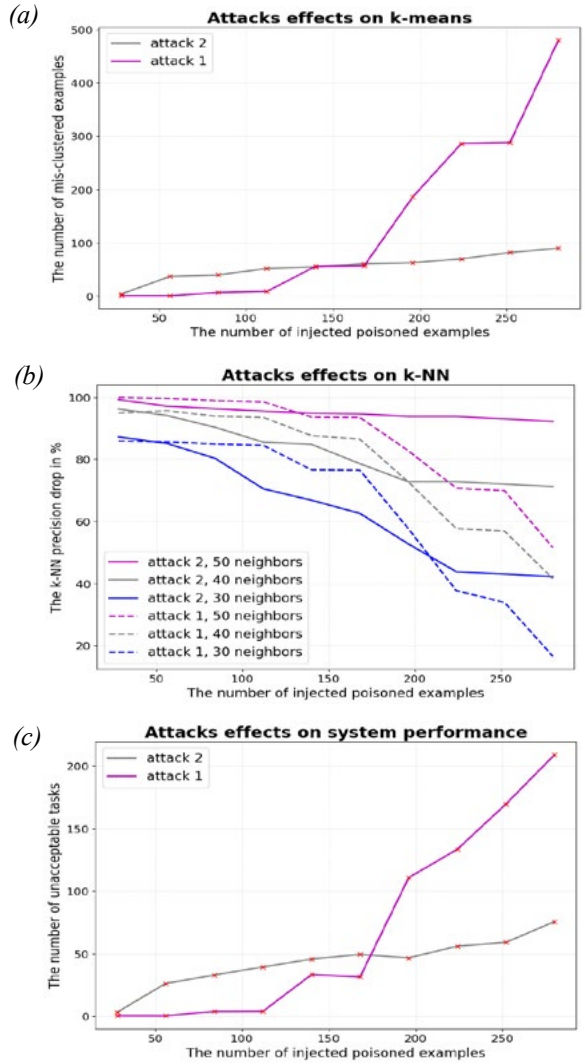


Fig.3. Attacks effects on (a) clustering, (b) prediction, and (c) system performance.

5. MITIGATION TECHNIQUES

Here below, we propose two attack mitigation methods based on anomaly detection. The first one (called *method 1*) is based on the Gaussian Mixture Model (GMM). Anomaly detection using GMM detects outliers by identifying the data points (generated tasks) in low-density regions [8]. The second method (called *method 2*) is based on the Silhouette Score test applied before GMM. Then, GMM is applied to the mean 3D data points of the considered Gaussian distributions. Both methods are semi-supervised learning algorithms where some history is needed to determine the parameters of GMM.

Method 1 uses GMM to fit the whole dataset with multiple Gaussian distributions. The task is decided to be anomalous if its position appears outside the high probability region which is determined by the optimization algorithm that maximizes the F1 score. (In our case, the threshold probability value equals 0.91.) The results presented in Fig. 4 show that method 1 is efficient for attack 2 as it has a higher recall[3]. It boosts the overall performance between 80% and 90% for different *k* values for the maximum F1 score. Here, the overall *performance boost* is the averaged performance of clustering, prediction, and system performance[4] after applying attack mitigation *method 1* compared to prior mitigation. On the other hand, method 1 is inefficient for attack 1. It suffers from low recall for some scenarios. This happens when the mean of the malicious source's generated data points are considered non-anomalous, or their high-probability region intersects with legitimate clusters.

Method 2 is used to solve the problem of the intersection of clusters. Here, the Silhouette score test is used prior to the anomaly detection to determine the optimum clusters' number. Based on these results, GMM is fitted to the model. The Q-Q plot for each distribution checks whether intersections between clusters occur. Because the sum of Gaussian distributions is also Gaussian, the

[3] Recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made.

[4] Clustering performance is the ratio of the number of erroneously classified tasks to the total number of tasks; Prediction performance is the ratio of the number of erroneously predicted tasks to the total number of predictions. The system performance is defined as the ratio between the number of erroneous allocations of tasks to 2C links and the total number of tasks.

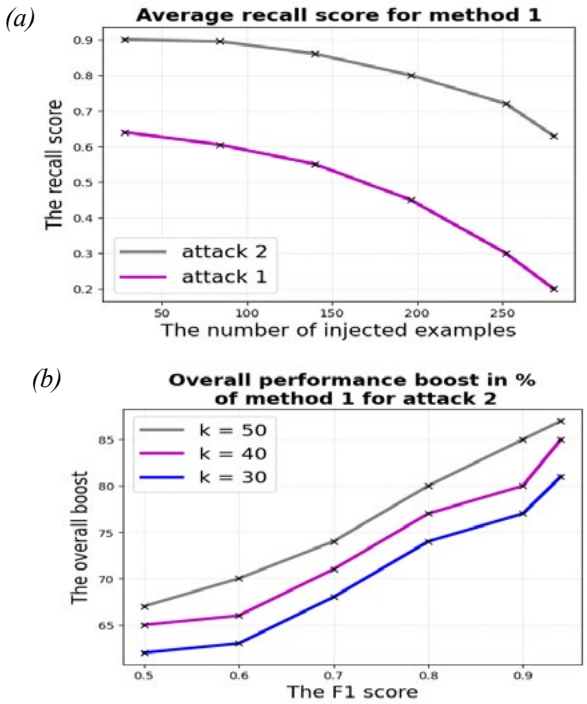


Fig.4. Method 1 performance (a) recall score for attacks 1 and 2, (b) performance boost for attack 2

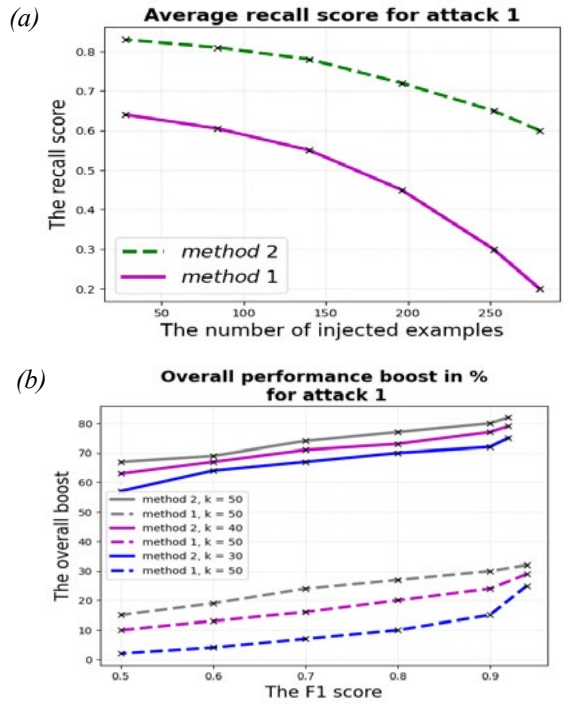


Fig.5. Method 2 compared to method 1 performance in case of attack 1, (a) recall score, (b) performance boost

separation might be done. Once the distributions are separated, anomaly detection on the distributions' means is performed. If the distribution mean is found outside the high probability region (defined for the optimum value maximizing F1 score and equal to 0.92), the whole distribution (cluster) is decided to be anomalous.

The results in Fig. 5 show that method 2 solves the intersection problem where the recall percentage is enhanced. Because the algorithm must discard one cluster, it suffers from both low precision and recall for scenarios where the malicious source distribution mean is missclassified. In addition, the algorithm fails to separate the clusters for lower Euclidean distance between the malicious and the nearest cluster as shown in Fig. 6.

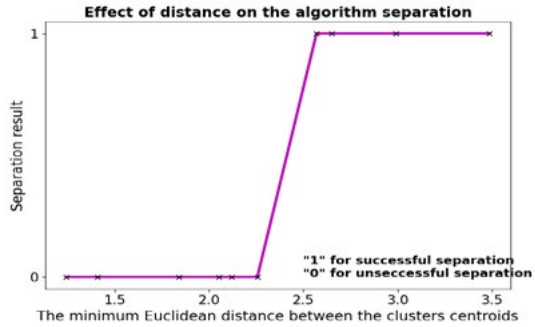


Fig. 6. Separation vs. clusters distancing

6. CONCLUSION

We considered 2C task classification based on *k*-means and *k*-NN to optimize their allocation to servers. We introduced two attack scenarios where the poisoning examples are grouped in the considered data region (attack 1) or spread (attack 2). The simulation results show that the number of correct task allocations is reduced by these attacks, which leads to the overload of some edge computing servers and the inability to perform the 2C tasks. To mitigate the attacks and improve the system's performance, we propose two anomaly detection methods. Method 1 is based on GMM anomaly detection for the whole data set, while method 2 is based on the Silhouette Score test prior to GMM, which is applied to the means of the identified clusters. Our simulation results show that method 1 detects the anomalies and improves the system performance if the clusters are well separated. However, if we face cluster intersections, the algorithm performs badly, and our method 2 is more efficient in solving this problem and detecting the poisoned data.

REFERENCES

[1] M. Hahsler, M. Piekenbrock, D. Doran. "dbscan: Fast density-based clustering with R." *J. of Stat. Software* 91 (2019): 1-30.

[2] J.A. Hartigan, M.A. Wong. "A k-means clustering algorithm." *Applied statistics* 28.1 (1979), 100-108.

[3] D.T. Larose, C.D. Larose. "k-nearest neighbor algorithm." (2014): 149-164.

[4] X. Zhang, X. Zhu, L. Lessard. "Online data poisoning attacks." *Learning for Dynamics and Control*. PMLR, 2020.

[5] P.J. Rousseeuw, "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis", *Comp. App. Math.,* Vol. 20, Nov. 1987, pp. 53–65.

[6] Z.C. Lipton, C. Elkan, B. Narayanaswamy. "Thresholding classifiers to maximize F1 score." *arXiv preprint arXiv:1402.1892* (2014).

[7] T. Balharith, F. Alhaidari. "Round robin scheduling algorithm in CPU and cloud computing: a review." *IEEE ICCAIS*, 2019.

[8] B. Yu, et.al. "A Network Traffic Anomaly Detection Method Based on Gaussian Mixture Model", *Electronics*, 2023, no. *12*(6), 1397, pp. 1-9

DEKODOWANIE KODÓW ITEROWANYCH Z UŻYCIEM SIECI NEURONOWEJ

DECODING OF ITERATED CODES WITH THE USE OF A NEURAL NETWORK

Oskar Fiedot[1]; Marek Blok[2]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, oskar.fiedot@gmail.com
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, marek.blok@pg.edu.pl

DOI: 10.15199/59.2023.4.40

Streszczenie: **Nadmiarowe kody iterowane są jedną z prostych metod pozyskiwania długich kodów korekcyjnych zapewniających dużą ochronę przed błędami. Jednocześnie, chociaż ich podstawowy iteracyjny dekoder jest prosty koncepcyjnie oraz łatwy w implementacji, to nie jest on rozwiązaniem optymalnym. Poszukując alternatywnych rozwiązań zaproponowano, przedstawioną w pracy, strukturę dekodera tego typu kodów wspomaganą przez sieci neuronowe. Zaproponowane rozwiązanie pozwala na wykrywanie oraz korekcję błędów w odbieranych ciągach.**

Abstract: **Redundant iterated codes are one of the simple methods of deriving long correction codes that provide high error protection. At the same time, although their basic iterative decoder is conceptually simple and easy to implement, it is not an optimal solution. Looking for alternative solutions, a neural network-assisted decoder structure for this type of codes was proposed. The solution presented in this paper allows the detection and correction of errors in the received sequences.**

Słowa kluczowe: **dekodowanie kodów iterowanych, detekcja i korekcja błędów, sieć neuronowa, syndrom**

Keywords: **iterated code decoding, error detection and correction, neural network, syndrome**

1. WSTĘP

W dzisiejszym świecie, w związku z coraz to nowszymi usługami sieciowymi, takimi jak na przykład strumieniowanie gier komputerowych lub aplikacji w czasie rzeczywistym, coraz szybciej rośnie ilość danych przesyłanych pomiędzy różnymi urządzeniami oraz wymagania na szybkość transmisji, jak i odporność na błędy powstające w trakcie transmisji. Wiele usług jest bardzo wrażliwych na tego typu błędy, które mogą w znacznym stopniu wpływać negatywnie na wygodę korzystania z tych usług [2]. Wymaga to projektowania takich systemów, które w jak najlepszym stopniu są w stanie wykrywać oraz poprawiać błędy powstające na różnych etapach przesyłania informacji z jednego urządzenia, na drugie.

Jednym ze sposobów na skuteczne wykrywanie oraz korekcję błędów jest korzystanie z liniowych kodów blokowych [2], które umożliwiają korekcję pewnej liczby błędów. Uzyskuje się to poprzez dodanie do przesyłanego ciągu bitów dodatkowych bitów kontroli parzystości. Wraz z długością kodu rosną jego potencjalne zdolności korekcyjne, jednak zwykle wiąże się to ze znacznym wzrostem złożoności jego dekodowania. Alternatywą jest złożenie jednego dłuższego kodu z kilku kodów składowych. Jednym z możliwych podejść do tego, jest stosowanie kodów iterowanych [2]. Ograniczeniem kodów iterowanych jest jednak to, że klasyczny, iteracyjny, dekoder tego typu kodów, mimo że bardzo prosty w implementacji, często nie jest w stanie poprawić wszystkich błędów, których poprawę gwarantują wysokie zdolności korekcyjne kodu iterowanego.

Coraz większą popularnością na świecie cieszą się również sieci neuronowe, będące implementacją sztucznej inteligencji, które próbuje się wykorzystywać wszędzie tam, gdzie mogą one wpłynąć pozytywnie na skuteczność projektowanych systemów. Znajdują one również zastosowanie w systemach informacyjnych, a przykładowe próby wytrenowania sieci neuronowych do dekodowania liniowych kodów blokowych zostały opisane w pracach [3] oraz [4].

W pracy przedstawiono strukturę dekodera kodów iterowanych, wykorzystującego możliwości sieci neuronowych do poprawy skuteczności dekodera w zakresie korekcji błędów oraz wyniki testów skuteczności tego rozwiązania.

2. TRENING I TESTY SKUTECZNOŚCI ZAPROPONOWANEJ SIECI

2.1. Struktura kodu iterowanego

Kod iterowany, dla którego poszukiwana była optymalna struktura sieci jest kodem podwójnie iterowanym, składającym się z kodów składowych zdolnych poprawić pojedynczy błąd. Do kodowania wierszy wybrano kod Hamminga o następującej macierzy generującej:

$$G_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \quad (1)$$

natomiast do kodowania kolumn, został wybrany kod o macierzy generującej:

$$G_2 = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \quad (2)$$

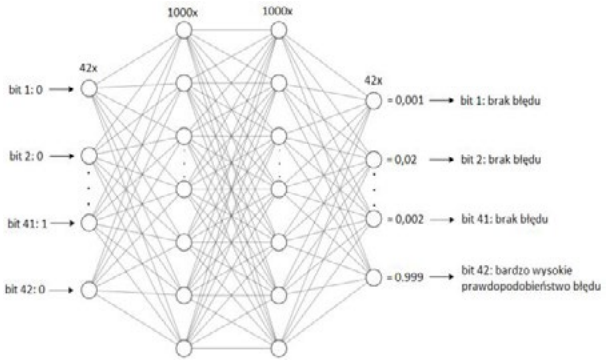
Kody te cechują kolejno 4 bity informacyjne i długość 7 oraz 3 bity informacyjne i długość 6, co oznacza, że wynikowy kod cechuje 12 bitów informacyjnych oraz łączna długość 42 bity, co daje 30 dodatkowych bitów kontroli parzystości. Zawartość informacyjna złożonego

kodu iterowanego jest równa 0,285, co oznacza, że zaledwie 28,5% wszystkich bitów w kodzie przenosi informacje bezpośrednio. Zdolności korekcyjne kodu gwarantują poprawę 4 błędnych bitów w ciągu odebranym.

2.2. Struktura sieci neuronowej

Zaproponowany dekoder opiera się o współpracę dwóch sieci neuronowych. To podejście jest wynikiem niewystarczająco dobrych rezultatów osiąganych przez dekoder oparty na działaniu pojedynczej sieci neuronowej. Dekodowanie początkowo było przeprowadzane tylko przez jedną z dwóch obecnie wykorzystywanych sieci, której zadaniem jest określanie czy poszczególne bity uległy przekłamaniu. Taka sieć posiada liczbę wejść oraz wyjść równą liczbie wszystkich bitów w kodzie iterowanym. W przypadku omawianego kodu iterowanego (rys. 1) oznacza to 42 neurony wejściowe oraz 42 neurony wyjściowe w sieci neuronowej. Sieć posiada również dwie warstwy ukryte, w której znajduje się po 1000 neuronów. Taka liczba neuronów sprawia, że sieć połączeń pomiędzy nimi jest bardzo skomplikowana, umożliwiając sieci neuronowej przetwarzanie oraz poznawanie złożonych relacji pomiędzy poszczególnymi bitami w ciągach kodowych oraz pozwala na szybkie osiąganie stosunkowo dobrych wyników predykcji sieci.

Każdy neuron na wyjściu oraz wejściu sieci odpowiada pojedynczemu bitowi w ciągu odebranym. Sieć na wejście przyjmuje ciąg zer i jedynek, będący ciągiem odebranym z wprowadzonymi błędami, natomiast zwraca na wyjściu ciąg wartości z zakresu od 0 do 1, reprezentujących rozmytą decyzję, czy poszczególne bity uległy przekłamaniu.



Rys. 1. Diagram sieci neuronowej oszacowującej czy poszczególne bity uległy przekłamaniu

Neurony w warstwach ukrytych mają funkcje aktywacji ReLU, która zeruje ujemne wartości, a wartości dodatnie przekazuje niezmienione. Warstwa wyjściowa wykorzystuje natomiast funkcję aktywacji Sigmoid. Funkcja ta zwraca zawsze wartości z zakresu od 0 do 1, przez co nadaje się do reprezentowania oszacowania binarnej decyzji i odpowiednio skaluje wartości na wyjściu sieci neuronowej. Jako optymalizator do treningu został wybrany algorytm Adama [1], będący metodą stochastycznego spadku gradientu. Jest to metoda wydajna obliczeniowo, o niewielkich wymaganiach pamięciowych i dobrze nadająca się do problemów, które są skomplikowane pod względem ilości danych oraz parametrów [1].

Trening opisywanej sieci poskutkował bardzo ciekawymi wynikami. Skuteczność sieci dla dekodowania ciągów, w których wystąpiły różne liczby błędów, została przedstawiona na rys. 2. Jak widać, sieć jest w stanie osiągnąć 100% skuteczności dla dekodowania ciągów odebranych, w których wystąpił jeden błąd, 52% skuteczności dla dekodowania ciągów z dwoma błędami, około 40% dla ciągów z trzema błędami oraz około 30% dla ciągów z czterema błędami. Spadająca skuteczność dla coraz większych liczb błędów nie jest zaskakująca, natomiast zaskakujące jest jak niską skuteczność osiąga sieć neuronowa podczas dekodowania ciągów, w których nie wystąpił żaden błąd. Skuteczność sieci w takim przypadku to niecałe 3%. Co ciekawe, sieć o wyżej opisanej strukturze, osiąga niemal 100% skuteczności dla dekodowania ciągów, w których wystąpił jeden błąd, nawet gdy trening został przeprowadzony z wyłączeniem ciągów, zawierających tylko pojedynczy błąd. Skuteczność sieci dla dekodowania bezbłędnych ciągów, nie wzrosła nawet wtedy, gdy znacznie zwiększono ilość danych treningowych, będących bezbłędnymi ciągami.

Do opracowania zaproponowanego rozwiązania kluczowe okazało się spostrzeżenie, że mimo zaledwie 33% skuteczności dla dekodowania ciągów z czterema błędami, prawie zawsze sieci udaje się poprawić przynajmniej jeden z czterech błędów (ponieważ skuteczność sieci dla poprawiania pojedynczych błędów to 100% dla danych testowych). Taka sytuacja została przedstawiona na rys. 3, gdzie sieci neuronowej nie udało się poprawić wszystkich czterech błędów w ciągu odebranym, jednak prawidłowo poprawiła dwa błędne bity z czterech (błędne bity: 12, 15, 26, 40; poprawione bity: 15, 40). Podobne sytuacje, w których sieć poprawia słusznie pewną część błędnych bitów, zachodzą również dla trzech oraz dwóch błędów w ciągu odebranym.

```
Testing the neural network on a set of messages containing 0 errors....
313/313 [==============================] - 1s 2ms/step - loss: 4.7084e-06 - accuracy: 0.0266
The accuracy achieved by the neural network: 2.66 %

Testing the neural network on a set of messages containing 1 errors....
313/313 [==============================] - 1s 2ms/step - loss: 1.0711e-05 - accuracy: 1.0000
The accuracy achieved by the neural network: 100.0 %

Testing the neural network on a set of messages containing 2 errors....
313/313 [==============================] - 1s 2ms/step - loss: 1.2015e-04 - accuracy: 0.5238
The accuracy achieved by the neural network: 52.38 %

Testing the neural network on a set of messages containing 3 errors....
313/313 [==============================] - 1s 2ms/step - loss: 0.0015 - accuracy: 0.3953
The accuracy achieved by the neural network: 39.53 %

Testing the neural network on a set of messages containing 4 errors....
313/313 [==============================] - 1s 2ms/step - loss: 0.0194 - accuracy: 0.3374
The accuracy achieved by the neural network: 33.74 %
```

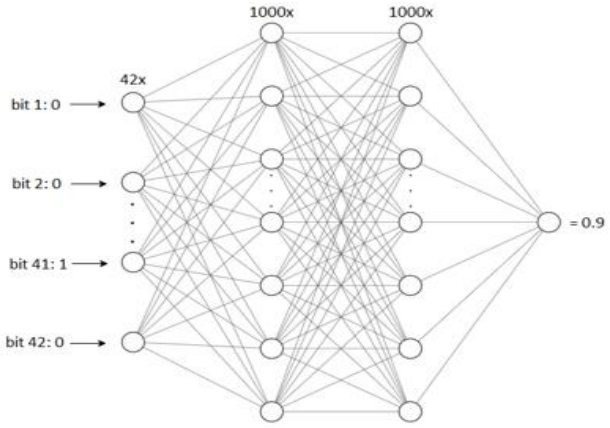
Rys. 2. Skuteczność sieci dla dekodowania ciągów, w których wystąpiły różne liczby błędów

```
Message received:
[0 1 0 0 1 0 1 1 1 1 1 0 1 1 0 0 0 1 0 0 1 1 0 1 1 1 1 1 0 1 1 0 1 1 0 0 0 1
 1 0 0 1 0]

The error vector returned by the neural network:
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 1 0 0]

Expected error vector:
[0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 1 0 0]
```

Rys. 3. Przypadek, w którym sieć neuronowa poprawnie wskazała dwa błędne bity z czterech (błędne bity: 12, 15, 26, 40; poprawione bity: 15, 40)

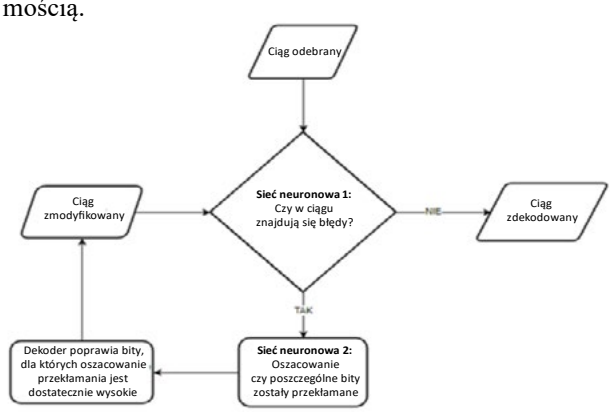
Aby naprawić problem niepoprawnego dekodowania bezbłędnych ciągów oraz wykorzystać zaletę zaproponowanej sieci neuronowej w postaci zdolności do po-

prawiania części błędnych bitów, zdecydowano się wytrenować dodatkową sieć neuronową, której zadaniem jest określanie czy w ciągu odebranym wystąpił co najmniej jeden błąd, bez wskazywania liczby błędów oraz ich pozycji. Ta dodatkowa sieć neuronowa ma taką samą strukturę oraz parametry treningu, jak wcześniej opisywana sieć, z tym wyjątkiem, że warstwa wyjściowa składa się z tylko jednego neuronu, którego wyjście odpowiada rozmytej decyzji dotyczącej wystąpienia błędów w ciągu odebranym (rys. 4).



Rys. 4. Diagram sieci neuronowej szacującej czy w ciągu odebranym wystąpił co najmniej jeden błąd

W ten sposób działanie dekodera zostało ostatecznie oparte na współpracy dwóch sieci neuronowych (rys. 5), podczas której, gdy na wejście dekodera trafia ciąg odebrany, jedna z sieci oszacowuje czy w ciągu pojawiły się błędy. Następnie, jeśli wartość otrzymana na wyjściu sieci jest bliska jedności, to ciąg odebrany przechodzi na wejście następnej sieci, która ocenia dla poszczególnych bitów czy zostały one przekłamane. Po poprawieniu wykrytych błędów wynikowy ciąg trafia z powrotem na wejście sieci, oszacowującej występowanie błędów w korygowanym ciągu. Jeżeli wartość oszacowania, że w ciągu znajdują się błędy jest dostatecznie niska, dekoder zwraca ciąg w bieżącej postaci, będący już zdekodowaną wiadomością.

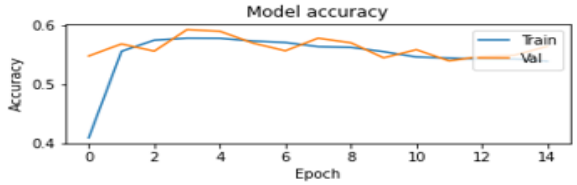


Rys. 5. Diagram przedstawiający algorytm postępowania w ramach zaproponowanego dekodera

2.3. Przebieg treningu

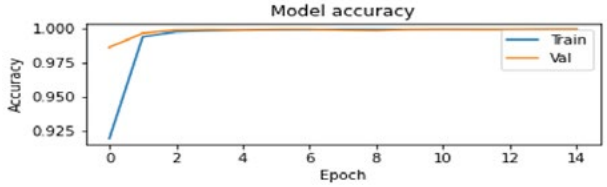
Do wytrenowania sieci dekodujących opisany kod iterowany konieczne było wygenerowanie odpowiednio dużego zbioru danych treningowych. Do treningu obydwu sieci wygenerowano po około milionie danych treningowych, w postaci ciągów kodowych z losowo wprowadzonymi błędami. Do tych danych dołączane są oczekiwane odpowiedzi sieci neuronowych w postaci wektorów błędów oraz jednobitowych informacji o tym czy w ciągu znajdują się błędy. Dane były generowane cyklicznie, co przy liczbie danych równej milionowi, oznacza, że każde słowo kodowe pojawiło się w zbiorze treningowym ponad 200 razy (ponieważ kod posiadający 12 bitów informacyjnych posiada $2^{12} = 4096$ słów kodowych). Takie podejście umożliwia osiągnięcie znacznie lepszych skuteczności treningu sieci neuronowych niż w przypadku danych wygenerowanych całkowicie losowo, gdy nie ma pewności co do tego, czy każde słowo kodowe pojawi się podczas treningu przynajmniej raz.

Najpierw przeprowadzono trening sieci służącej do identyfikacji przekłamania poszczególnych bitów. Dane zostały podzielone na dane treningowe (70%), walidacyjne (15%) oraz testowe (15%). Trening składał się z 15 epok. Na rys. 6 przedstawiono wykres skuteczności wytrenowanej sieci neuronowej. Jak widać, już podczas pierwszej epoki udało się uzyskać skuteczność sieci neuronowej na poziomie ponad 50%. Podczas kolejnych epok skuteczność sieci utrzymywała się na stosunkowo równym poziomie. Więcej informacji na temat skuteczności sieci, może dostarczyć rys. 2. Zważywszy na sposób działania zaproponowanego dekodera, najważniejsza jest skuteczność sieci dla dekodowania ciągów z pojedynczym błędem, ponieważ można ten parametr rozumieć też jako szansę na to, że sieć poprawi przynajmniej jeden błąd ze wszystkich obecnych w ciągu (o ile liczba błędów nie przekracza zdolności korekcyjnych kodu iterowanego). Sieć w tym przypadku osiąga niemal 100% skuteczności, co jest wynikiem jak najbardziej zadowalającym.



Rys. 6. Wykres zmian dokładności sieci służącej do wykrywania błędnych bitów dla kolejnych epok treningu

W analogiczny sposób podzielono dane służące do treningu sieci neuronowej, której zadaniem jest określanie czy w ciągu odebranym pojawił się przynajmniej jeden błąd. Parametry treningu oraz struktura sieci również pozostają takie same z wyjątkiem liczby neuronów w warstwie wyjściowej.



Rys. 7. Wykres dokładności sieci do wykrywania, czy w wiadomości wystąpiły, dla kolejnych epok treningu

Jak widać na rys. 7, już przy pierwszej epoce udało się uzyskać bardzo wysoką skuteczność sieci na poziomie

około 98%. Skuteczność sieci utrzymywała się na równym poziomie przez cały czas trwania treningu podobnie jak w przypadku poprzedniej sieci neuronowej. Na rys. 8 można zauważyć, że wytrenowana sieć neuronowa osiąga skuteczność na poziomie ponad 99% dla ciągów z losową liczbą błędów (w zakresie zdolności korekcyjnych kodu) na losowych pozycjach, co jest znakomitym wynikiem.

```
4688/4688 [==============================] - 10s 2ms/step - loss: 0.0037 - accuracy: 0.9995
The accuracy achieved by the neural network: 99.95 %
```

Rys. 8. Skuteczność sieci neuronowej służącej do określania, czy w wiadomości wystąpiły błędy

W ten sposób, możliwe było wytrenowanie dwóch sieci neuronowych, które mogą zostać wykorzystane w zaproponowanym dekoderze, przeprowadzającym dekodowanie wiadomości w oparciu o współpracę tych dwóch sieci neuronowych. Na rys. 9 przedstawiono przykład ciągu zawierającego 4 błędy poprawnie zdekodowanego z użyciem zaproponowanego dekodera.

```
Message received:
[0 1 0 0 0 0 1 1 0 1 1 0 1 0 0 1 1 0 0 0 1 1 1 0 0 0 1 0 0 0 0 0 0 0 0 1 1
 0 0 0 1 1]

Error vector:
[0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0]

Message corrected by a neural network-based decoder:
[0 1 1 1 0 0 1 1 0 1 1 0 1 0 0 1 1 1 0 0 1 1 1 0 0 0 1 1 0 0 0 0 0 0 0 1 1
 0 0 0 1 1]
```

Rys. 9. Przykład poprawnie zdekodowanej wiadomości

3. Porównanie z dekoderem klasycznym

Z punktu widzenia użyteczności dekodera opartego na sieciach neuronowych, istotne jest również porównanie jego skuteczności ze skutecznością dekodera klasycznego. Wyniki porównania dla danych z losową liczbą błędów przedstawiono na rys. 10. Jak można zauważyć, obydwa dekodery osiągnęły bardzo wysokie i podobne skuteczności na zadanym zbiorze danych testowych. Przewagę dekodera opartego na sieciach neuronowych, można zauważyć przy porównaniu skuteczności dekodowania ciągów, w których wystąpiły 4 błędy. To porównanie zostało przedstawione na rys. 11. W tym przypadku, dekoder oparty na sieciach neuronowych osiągnął skutecznością lepszą o niecałe 5 punktów procentowych. Jednak w przypadku liczb błędów mniejszych niż 4 dekoder klasyczny osiągał zawsze skuteczność równą 100%, natomiast dekoder oparty na sieciach neuronowych ponad 99%, co, mimo że jest to minimalna przewaga, daje jednak pewność, że dekoder klasyczny poprawnie zdekoduje ciągi, w których wystąpiły te liczby błędów, podczas gdy tej pewności nie ma przy stosowaniu dekodera, opartego na sieciach neuronowych. Na podstawie tych wyników, można wysnuć wniosek, że dekoder, oparty na sieciach neuronowych, w większości przypadków nie gwarantuje 100% skuteczności dekodowania, jednak może zapewnić skuteczność na poziomie około 99% dla każdej możliwej do poprawy liczby błędów, w przeciwieństwie do dekodera klasycznego, który choć zapewnia 100% skuteczności dla dekodowania ciągów, w których wystąpiło do trzech błędów, cechuje się znacznie niższą skutecznością (choć nadal wysoką) podczas dekodowania ciągów, w których wystąpiły cztery błędy, mimo że zdolności korekcyjne kodu iterowanego powinny zapewniać poprawę takiej liczby błędów.

```
Comparing the effectiveness of decoders... ———————————————— 100% 0:00:00
Efficiency of a neural network-based decoder: 99.68 %
Efficiency of the classical decoder: 98.98 %
```

Rys. 10. Porównanie skuteczności dekodowania ciągów, w których wystąpiła losowa liczba błędów

```
Comparing the effectiveness of decoders... ———————————————— 100% 0:00:00
Efficiency of a neural network-based decoder: 98.82 %
Efficiency of the classical decoder: 94.04 %
```

Rys. 11. Porównanie skuteczności dekodowania ciągów, w których wystąpiły 4 błędy

4. Podsumowanie

Zaproponowany dekoder złożony z dwóch sieci neuronowych oraz metoda jego treningu pozwala na uzyskanie bardzo wysokiej skuteczności dekodowania i zarazem prostej implementacji. Udało się wykazać, że wytrenowanie sieci neuronowych do dekodowania kodów iterowanych jest możliwe, choć może wymagać bardzo dużej ilości danych treningowych, a dekoder oparty na sieciach neuronowych może być bardzo dobrą alternatywą dla dekodera klasycznego. Trening sieci neuronowych do dekodowania kodów iterowanych może jednak sprawiać pewne problemy. Tak wysoką skuteczność sieci udało się uzyskać dzięki dużej objętości zbioru danych treningowych, który był kilkukrotnie razy większy, niż zbiór wszystkich możliwych słów kodowych kodu iterowanego, dla którego trenowane były sieci. Umożliwiło to wykorzystanie każdego słowa kodowego w trakcie treningu bardzo wiele razy. W przypadku dłuższych kodów iterowanych, zbiór wszystkich słów kodowych jest tak ogromny, że wygenerowanie kilkukrotnie większego zbioru treningowego może być bardzo trudne lub praktycznie nieopłacalne. Oznaczałby to również bardzo długi trening sieci neuronowych i mogłoby wpłynąć na pogorszenie skuteczności. W dalszej perspektywie warto rozważyć połączenie w ramach proponowanej koncepcji dekodera klasycznego z dekoderem neuronowym wykorzystując każdy w zakresie, w którym jest on skuteczniejszy. Pozwoliłoby to na wykorzystanie dużej skuteczności klasycznego dekodera na mniejszej liczby błędów oraz ograniczyć problem stawiany przed siecią neuronową pozwalając na jej uproszczenie.

LITERATURA

[1] Kingma P. Diederik, Ba Jimmy. 2014. „Adam: A method for stochastic optimization".

[2] Fiedot Oskar. 2023. „Dekodowanie kodów iterowanych z użyciem sieci neuronowej". *Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki.*

[3] Abdelbaki Hisham, Gelenbe Erol, El-Khamy Said Esmail. 1999. "Random neural network decoder for error correcting codes". *International Joint Conference on Neural Networks*: 3241-3245.

[4] El-Khamy S. E, Youssef E. A., Abdou H. M., 1995, "Soft decision decoding of block codes using artificial neural network", *Proceedings IEEE Symposium on Computers and Communications*: 234-240.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

SmokeFinder - wspomagany algorytmami sztucznej inteligencji system wizyjnej detekcji dymu

SmokeFinder - a visual smoke detection system supported by artificial intelligence algorithms

Konrad Zuchniak[1]; Eryk Kosk[2]; Piotr Jaglarz[3];

[1] Neuralbit Technologies, Kraków, konrad.zuchniak@neuralbit.pl
[2] Neuralbit Technologies, Kraków, eryk.kosk@neuralbit.pl
[3] Neuralbit Technologies, Kraków, piotr.jaglarz@neuralbit.pl

DOI: 10.15199/59.2023.4.41

Streszczenie: SmokeFinder to system umożliwiający stałe analizowanie obrazu pochodzącego z kamer obserwacyjnych w celu wyszukiwania dymu w lesie. W przypadku wykrycia nawet niewielkich słupków dymu do operatora systemu wysyłane są ostrzeżenia. Zaawansowane algorytmy uczenia maszynowego analizują dane zebrane z kamery obserwacyjnej w czasie rzeczywistym. W przypadku wykrycia dymu ustalana jest jego lokalizacja. Dane z wielu kamer są łączone w celu usprawnienia akcji gaśniczej. W przypadku potwierdzenia zagrożenia automatycznie zaalarmowane zostaną odpowiednie lokalne jednostki straży pożarnej.

Abstract: SmokeFinder is a system that allows to constantly analyze the image from surveillance cameras in order to search for smoke in the forest. If even small pillars of smoke are detected, alerts are sent to the system operator. Advanced machine learning algorithms analyze the data collected from the observation camera in real time. If smoke is detected, its location is determined. Data from multiple cameras are combined to improve firefighting. If the threat is confirmed, the relevant local fire departments will be automatically alerted.

Słowa kluczowe: przetwarzanie brzegowe, strumieniowanie wideo, uczenie maszynowe, widzenie maszynowe

Keywords: edge computing, machine learning, machine vision, video streaming

1. WSTĘP

System SmokeFinder umożliwia rozszerzenie funkcjonalności już istniejącego systemu monitoringu wizyjnego lasu o automatyczną detekcję dymu, jak również w połączeniu z odpowiednią kamerą stanowić może kompletny system. Algorytm detekcji dymu opiera się o sztuczne sieci neuronowe – zastosowano konwolucyjne sieci neuronowe z rozszerzonym wejściem na dane kontekstowe.

System składa się z jednostki przetwarzania brzegowego (ang: edge processing) (dalej EPU, od ang: edge processing unit) zamontowanego zaraz przy kamerze obserwacyjnej. Jest to specjalnie zaprojektowany sprzęt zainstalowany na wieży obserwacyjnej. Takie rozwiązanie pozwala na dostęp do nieprzetworzonego sygnału wideo, który w przypadku analizy w chmurze czy innej lokalizacji niż kamera wymagałby stratnej kompresji w celu transmisji danych. Należy zwrócić uwagę, że często wieże obserwacyjne instalowane są w trudno dostępnych miejscach, gdzie mamy słaby zasięg sieci komórkowych. Oczywiście można instalować prywatną infrastrukturę łączności jednak jest to niewspółmiernie drogie. Nasze podejście umożliwia jednocześnie wysyłanie skompresowanego strumienia wideo w jakości dostosowanej do możliwości łącza do punktu alarmowo-dyspozycyjnego, w którym znajduje się obserwator kierujący ewentualną akcją gaśniczą przy jednoczesnej analizie obrazu nieskompresowanego, w najwyższej możliwej jakości.

Drugą częścią rozwiązania jest oprogramowanie chmurowe sterujące pracą wielu EPU zlokalizowanych na różnych wieżach obserwacyjnych. Ta część systemu odpowiada za wyświetlanie odpowiednich informacji końcowemu użytkownikowi, synchronizację pracy pomiędzy EPU, sterowanie EPU przez użytkownika czy zarządzanie systemem wyświetlania alarmów i wysyłania notyfikacji o wykryciu dymu. Działanie kamer jest ze sobą zsynchronizowane, punkt przecięcia kierunków patrzenia jest zaznaczony na mapie i lokalizacja tego przecięcia jest estymowana. Kiedy jedna z kamer wykryje dym, pozostałe kamery przechodzą w tryb szukania tego dymu. Gdy go znajdą, wszystkie zwrócone są w stronę zidentyfikowanego dymu, a punkt przecięcia ich osi obserwacji określa lokalizację pożaru. Dodatkowo ten mechanizm pełni też funkcję automatycznego wykluczania fałszywych alarmów. Kiedy kamera A wykryje potencjalny dym, kamera B próbuje wykryć w obszarze wspólnym dla A i B dym. Jeśli go nie znajdzie, prawdopodobieństwo detekcji faktycznego dymu ulega zmniejszeniu. Jeśli kamera B dym znajdzie, prawdopodobieństwo detekcji ulega znacznemu wzrostowi.

Urządzenia mogą pracować w dwóch trybach transmisji informacji. W pierwszym trybie, opisanym powyżej, transmitowany jest obraz z kamer na żywo. Służy on przede wszystkim realizacji zadania wspierania obserwatora. W sytuacji średniego i wysokiego zagrożenia pożarowego, większość jurysdykcji wymaga obecności ludzkiego obserwatora, a nasz system działa jako rozwiązanie wspierające przy nużącej obserwacji obrazu w wykrywaniu dymu. W drugim trybie pracy wysyłane są jedynie poszczególne klatki zawierające dym wraz z wyliczoną lokalizacją. Ten tryb pracy zakłada w pełni autonomiczne funkcjonowanie dedykowane niskim zagrożeniom pożarowym, kiedy ludzka obserwacja nie jest wymagana - niskie ryzyko pożaru, bardzo wolne tempo rozprzestrzeniania się ognia, dlatego nawet relatywnie opóźniona reakcja gaśnicza bę-

dzie wystarczająca. W tym trybie odpowiednio skonfigurowane notyfikacje będą wysyłane np. na aplikację mobilną pracowników nadleśnictwa, zawierające zdjęcie z potencjalnym dymem oraz lokalizacją. Użytkownik w tym scenariuszu potwierdza lub odrzuca alarm.

We wszystkich scenariuszach użytkowania wykorzystujemy funkcję feedbacku pochodzącego od specjalisty. Kiedy użytkownik odrzuci lub potwierdzi alarm, informacje te automatycznie rozszerzają zbiór danych i umożliwiają dostrajanie do lokalnych warunków i udoskonalanie modelu detekcji opartego o uczenie maszynowe.

Projekt pod tytułem "SmokeFinder - wspierany algorytmami sztucznej inteligencji, uniwersalny i łatwy w instalacji system wizyjnej detekcji dymu" był współfinansowany ze środków Funduszy Europejskich w ramach Programu Operacyjnego Inteligentny Rozwój 2014-2020.

2. AKWIZYCJA DANYCH

W ramach współpracy z Polskim Instytutem Badawczym Leśnictwa pozyskaliśmy nagrania z pożarów leśnych. W wybranych punktach obserwacyjnych zainstalowaliśmy rejestratory, które gromadziły nagrania wideo. Następnie zidentyfikowaliśmy klatki, na których widoczny jest dym i ustaliliśmy lokalizację dymu na zdjęciach. Dodatkowo jako źródło danych wykorzystaliśmy dwa mniejsze publicznie dostępne zbiory danych zawierające zdjęcia i filmy z dymem.

1. Zestaw danych dotyczących wykrywania dymu utworzony w lipcu 2012 r. na Uniwersytecie w Salerno. Składa się z 149 filmów, z których każdy trwa około 15 minut, co daje ponad 35 godzin nagrania [1].
2. Baza danych dymu zebrana przez Centrum Badań Pożarów Leśnych (Wydział Elektrotechniki, Inżynierii Mechanicznej i Architektury Morskiej Uniwersytetu w Splicie) zawiera kolekcję niesegmentowanych obrazów dymu z pożarów lasów, sfotografowanych zarówno z ziemi, jak i z powietrza. [2]

3. OPRACOWANIE SPOSOBU ETYKIETOWANIA ZDJĘĆ

Korzystając z autorskiego oprogramowania do etykietowania zdjęć (na podstawie otwartego programu LabelImg [3]), zidentyfikowaliśmy ramki, w których widoczny jest dym oraz ustaliliśmy lokalizację dymu na zdjęciach. Zastosowano annotację obrazu typu bounding box oraz format annotacji obrazu Pascal VOC [4]. Postanowiliśmy potraktować ten problem przede wszystkim jako problem klasyfikacji, a nie problem wykrywania obiektów. Decyzja podyktowana była głównie większą szybkością algorytmów klasyfikacji, ale napotkaliśmy również podstawowy problem z poprawną annotacją dymu. Nie jest to przedmiot o ostrych krawędziach, dlatego obwiednie narysowane przez osoby etykietujące miały stosunkowo duży błąd. Nie jest możliwe uchwycenie idealnych krawędzi dymu, tak jak można dokładnie zaznaczyć osobę czy samochód. Bez problemu można generować klatki, na których z pewnością dym się znajduje, jednak przygotowanie zbioru danych pod problem detekcji obiektów jest obarczone pewnym rozmyciem co negatywnie wpływało na jakość uzyskiwanego detektora obiektów.

Aby rozwiązać ten problem, opracowaliśmy inne podejście. Mając jedną ramkę określającą położenie dymu, możemy wygenerować wiele wycinków obrazu, które dzielą część powierzchni z annotacją. W ten sposób z jednego zdjęcia możemy wygenerować wiele próbek; przycinamy różne fragmenty tego samego zdjęcia, na którym widać dym, dzięki czemu w ten sposób udało się dokonać dodatkowej augmentacji zbioru danych.

Na wygenerowane wycinki obrazków nałożyliśmy kilka warunków brzegowych. Określono maksymalny i minimalny procent powierzchni zadymienia. Zdefiniowano maksymalne i minimalne wartości procentowe obszaru annotacji, które muszą być uwzględnione w wycinku. Określono maksymalny i minimalny stosunek długiego boku do krótszego boku, a także maksymalną i minimalną powierzchnię wybranej próbki. Następnie wybrane fragmenty są skalowane do docelowego rozmiaru, np. 224×224 piksele. Dodatkową korzyścią jest to, że takie podejście pozwala wygenerować końcowy zestaw danych treningowych z różnymi rozkładami modyfikując warunki brzegowe generowania wycinków i modyfikując filtrowanie annotacji, które chcemy uwzględnić w danych źródłowych. W ramach prac zostały opracowane skrypty do automatycznego generowania finalnych zbiorów danych treningowych o różnych parametrach. Dodatkowo opracowano mechanizmy umożliwiające wygodne zarządzanie zgromadzonym zbiorem danych.

4. ROZSZERZENIE DANYCH

Od samego początku zbierania danych mieliśmy na uwadze plan wykorzystania danych kontekstowych. Zbierane kontekstowe informacje podzieliliśmy na dwie główne kategorie, które rozróżniają dane na sposób ich pozyskiwania oraz wartości fizyczne jakie reprezentują. Dane podzieliliśmy na kategorie meteorologiczne oraz lokalizacyjne. Poniżej przedstawiamy skrótowe opisy tych danych oraz sposobów ich pozyskiwania.

1. **Dane meteorologiczne** - we wszystkich lokalizacjach, w których prowadziliśmy akwizycję danych zbieraliśmy również dodatkowe dane meteorologiczne. W celu zbierania tych danych wykorzystaliśmy API OpenWeatherMap [5]. Zbieraliśmy informacje takie jak: temperatura, wilgotność, nasłonecznienie, ciśnienie atmosferyczne, zachmurzenie, kierunek i szybkość wiatru. Dane były zbierane w rozdzielczości czasowej jednej godziny. Informacje te zapisywaliśmy w plikach CSV, co następnie ułatwiło parowanie konkretnego rekordu z konkretnym nagraniem.
2. **Dane lokalizacyjne** - na podstawie przeprowadzonych rozmów z ekspertami domenowymi, osobami zajmującymi się zawodowo zabezpieczaniem przeciwpożarowym lasów i obserwacją prewencyjną doszliśmy do wniosku, że dodatkowym bardzo istotnym wektorem informacji są dane o położeniu przestrzennym kamery, położeniu słońca i względnym położeniu słońca względem kamery. W zależności od źródeł nagrań opracowane zostały nieco inne techniki akwizycji i standaryzacji danych określających aktualne położenie kamery. Finalnie dla znaczącej większości zebranych klatek dysponujemy orientacją przestrzenną definiowaną jako kąt wertykalny (kąt jaki tworzy się pomiędzy linią obserwacji, a kierunkiem północ), kąt nachylenia (kąt w

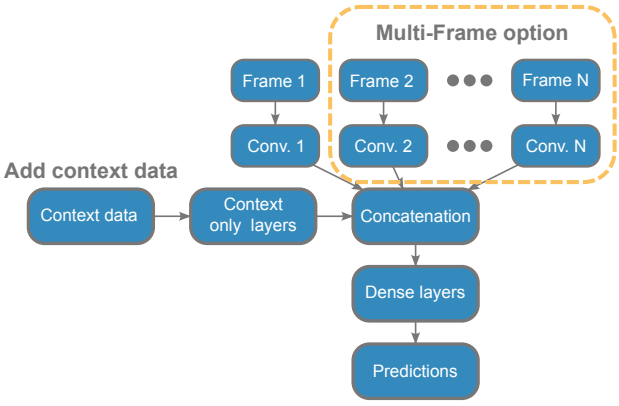
górę lub dół jaki tworzy się między linią obserwacji a linią przecinającą horyzont) wartość ta głównie jest bliska zeru, co jest spowodowane prowadzeniem obserwacji głównie w stronę horyzontu, zatem z nachyleniem bliskim zerowemu. Informację uzupełniającą stanowi aktualna wartość zoomu optycznego zastosowanego przez kamerę. Dodatkowo dla każdej współrzędnej geograficznej, dla zadanej daty i godziny możemy obliczyć kątowe pozorne położenie słońca na niebie, implementując specjalnie przeznaczone do tego algorytmy [6]. Mając te dwa położenia jesteśmy w stanie wyliczyć również względne położenie słońca względem linii obserwacji. Celem tego podejścia jest umożliwienie algorytmom rozróżnienia jaka była pozycja słońca dla każdej klatki, innymi słowy w ten sposób algorytm będzie "wiedział" czy zdjęcie było robione pod słońce czy nie, a to znacząco wpływa na wygląd zdjęcia i wygląd zarejestrowanego dymu. Dane lokalizacyjne rozszerzyliśmy też o współrzędne geograficzne.

5. OPRACOWANE MODELE

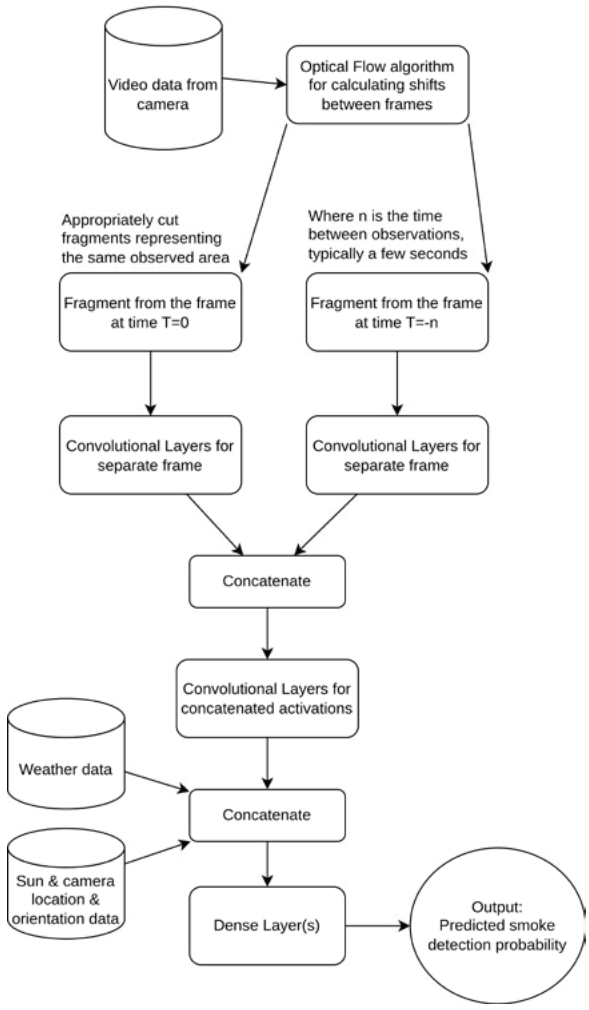
Opracowano i porównano ze sobą dwie główne koncepcje, które zostały wyłonione poprzez przeprowadzenie koncepcyjnego podziału sformułowania problemu. Założyliśmy, że nasz algorytm może działać zarówno jako klasyfikator binarny, który na wejściu otrzymuje obraz, a na wyjściu zwraca prawdopodobieństwa obecności dymu na analizowanym fragmencie. W tym podejściu wysokorozdzielcza klatka cięta jest na niewielkie fragmenty, które analizowane są niezależnie. W ten sposób otrzymujemy mapę prawdopodobieństw detekcji dymu na całym zdjęciu. Drugim podejściem jest detekcja obiektów. Na wejście algorytmu podajemy wtedy całe zdjęcie, a otrzymujemy propozycję współrzędnych prostokątów, na których zlokalizowany jest dym wraz z prawdopodobieństwem określających pewność modelu co do takich lokalizacji. Plusem tego podejścia jest procesowanie całego zdjęcia w jednym momencie, jednak wiąże się to ze znacznie dłuższym czasem inferencji i zwiększoną wielkością modelu co może sprawiać problemy przy próbach uruchomiania takiego modelu w urządzeniu offline. Dodatkowym problemem przy sformułowaniu zagadnienia jako detekcja obiektu było rozmycie dymu i trudności z wygenerowaniem dokładnych annotacji. Trudno jest dokładnie zaznaczyć lokalizację dymu na zdjęciu, ponieważ jego krawędzie nie są ostre.

Z kilku powodów klasyfikacja okazała się skuteczniejszą i bardziej adekwatną metodą. Po pierwsze zbiór danych, gdzie etykiety oznaczają: "fragment zawiera dym" oraz "fragment nie zawiera dymu" mogą być generowane ze znacznie większą precyzją względem rysowania prostokątów określających dokładne położenie dymu na analizowanej klatce, ponieważ dokładne oznaczenie gdzie dym się znajduje jest trudne, przez rozmyty charakter samego dymu. Po drugie modele klasyfikacji są prostsze i bardziej elastyczne, umożliwiając np. dodanie danych kontekstowych.

Dodatkowo mechanizm klasyfikacji łatwiej rozszerzyć o adaptatywną heurystykę i rozbić proces analizowania klatki na kilka etapów, co w wypadku detekcji obiektów było by w zasadzie niemożliwe. Przykładowo, w pierwszej fazie zdjęcie możemy podzielić np. na 12 dużych kwadratowych fragmentów i na nich przeprowadzić inferencję modelu w celu detekcji dymu. Następnie, jeżeli pewien fragment zwróci prawdopodobieństwo powyżej pewnej niewielkiej wartości progowej, możemy zwiększyć czułość algorytmu poprzez podział "podejrzanego" fragmentu na mniejsze fragmenty i przeanalizowanie ich osobno. Drugą oś podziału stanowi podejście jednoklatkowe względem podejścia wieloklatkowego. W pierwszym wariancie wejściem dla algorytmu jest jedna klatka pochodząca z kamer wideo. W podejściu drugim wejściem dla algorytmu jest sekwencja dwóch lub więcej klatek. Podejście wieloklatkowe umożliwia analizę temporalnych cech dymu, obserwacje fluktuacji i dyfuzji dymu w czasie, co przekłada się na lepsze odróżnianie dymu od podobnych obiektów. Wykorzystując dane kontekstowe mogliśmy zmodyfikować modele uczenia maszynowego, Wykorzystywaliśmy dane meteorologiczne oraz dane lokalizacyjne. Stosowaliśmy tutaj względnie standardowy preprocesing danych polegający na standaryzowaniu wartości cech względem zbioru trenującego. Dane dotyczące wiatru zostały skonwertowane z szybkości i kierunku wiatru na rozbitą szybkość wiatru w osi X i osi Y. Ze względu na swoją czasową charakterystykę danych meteorologicznych, rozważaliśmy wykorzystanie tylko najaktualniejszych punktów czasowych oraz szeregów czasowych reprezentujących również historyczne wartości. Zatem jako część odpowiedzialną za analizę tylko danych meteorologicznych testowaliśmy zarówno warstwy głębokie jak i rekurencyjne, takie jak warstwy LSTM, które są zdolne do wychwytywania również czasowych wzorców. Z testów zaobserwowaliśmy jednak, że warstwy głębokie analizujące ostatni punkt czasowy lub kilka ostatnich punktów czasowych dają podobny uzysk względem rekurencyjnych analiz długookresowych. Zapewne wynika to z faktu, że dla modelu detekcji dymu na obrazie najważniejsza jest informacja czy np. aktualnie występuje mgła, a nie np. czy dwa dni temu padał deszcz (to też są istotne dane, jednak w innych problemach, np. przy estymacji ogólnego zagrożenia pożarowego). Wykorzystanie danych lokalizacyjnych jest prostszym zagadnieniem, ponieważ są to typowe dane tabelaryczne. Kontekst czasowy jest tutaj nieistotny, dlatego wykorzystujemy tylko zestaw danych określających ostatni pomiar. Dane kontekstowe zostały dodane poprzez utworzenie dodatkowego wejścia w modelach, które przetwarzały te dane, a następnie łączyły je z wektorem cech generowanych przez warstwy konwolucyjne analizujące obraz. Rysunek 1 przedstawia schematyczne uwzględnienie danych kontekstowych w modelach. Rysunek 2 przedstawia schemat blokowy pełnej heurystyki detekcji dymu.



Rysunek 1: Schemat wprowadzenia danych kontekstowych dla modelu klasyfikacji jedno- (i opcjonalnie wielo-) klatkowego. Dodatkowe dane kontekstowe są wprowadzane do modelu klasyfikacji, następnie analizowane są przez dedykowane im warstwy, finalnie są łączone z wektorem cech generowanych przez część konwolucyjną modelu i na podstawie tych połączonych informacji tworzona jest finalna predykcja klasyfikacji dymu.



Rysunek 2: schemat architektury modelu uczenia maszynowego.

6. PODSUMOWANIE

W niniejszym opisie przedstawiliśmy pokrótce mechanizm działania systemu SmokeFinder, który wykorzystuje uczenie maszynowe celem umożliwienia najszybszej możliwej detekcji dzikich pożarów i tym samym umożliwiania natychmiastowej reakcji służb. Obraz z kamer obserwacyjnych jest stale analizowany w poszukiwaniu dymu. Kiedy urządzenie wykryje alarm, operator systemu od razu jest o tym fakcie informowany. System integruje również pracę wielu kamer. Kiedy jedna kamera wykryje dym, inne kamery również próbują zlokalizować ten obszar. Dodatkowo, podczas działania w trybie automatycznym system sam będzie poszukiwać dymu, a kiedy taki wykryje prześle odpowiednią notyfikację zawierającą podejrzany fragment nagrania oraz estymowaną lokalizację pożaru. Tak skonstruowany scenariusz działania umożliwia znaczne skrócenie czasu od zaprószenia ognia do przybycia służb gaśniczych, znacznie redukując tym samym szkody.

LITERATURA

[1] Lascio Rosario Di, Greco Antonio, Saggese Alessia, Vento Mario. 2014. "Improving fire detection reliability by a combination of videoanalytics". International Conference Image Analysis and Recognition, 477–484.

[2] Strona internetowa zbioru danych "Croatioa Smoke Dataset", http://wildfire.fesb.hr/index.php?option=com_content&view=article&id=49&Itemid=54

[3] Strona internetowa repozytorium programu LabelImg, https://github.com/heartexlabs/labelImg

[4] Everingham Mark, Van Gool Luc, Williams Christopher KI, Winn John, Zisserman Andrew. 2010. "The pascal visual object classes (voc) challenge". International journal of computer vision, 303–338.

[5] Strona internetowa dokumentacji Open Weather , https://openweathermap.org/api

[6] Reda Ibrahim, Andreas Afshin. 2008. "Solar position algorithm for solar radiation applications (revised)". ational Renewable Energy Lab.(NREL), Golden, CO (United States).

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

ANALIZA WPŁYWU ATAKÓW ZATRUWAJĄCYCH NA DETEKCJĘ ZASOBÓW RADIOWYCH BAZUJĄCĄ NA UCZENIU FEDERACYJNYM

ANALYSIS OF THE IMPACT OF POISONING ATTACKS ON FEDERATED LEARNING-BASED RADIO RESOURCE SENSING

Małgorzata Wasilewska[1,2]; Hanna Bogucka[1,2]

[1]Instytut Radiokomunikacji, Politechnika Poznańska
malgorzata.wasilewska@put.poznan.pl
hanna.bogucka@put.poznan.pl
[2]RIMEDO Labs, Polska

DOI: 10.15199/59.2023.4.42

Streszczenie: W niniejszym artykule przedstawiono wyniki badań i analizy wpływu ataków zatruwających odwracających etykiety (ang. label-flipping) na uczenie federacyjne w zastosowaniu dla detekcji zajętości zasobów radiowych. Badania przeprowadzono zarówno dla ataków skoordynowanych jak i losowych, przy zmiennym stosunku liczby użytkowników atakujących do liczby użytkowników uczciwych oraz różnym stopniu agresywności i czasie trwania ataków. Badania skupiają się na porównaniu skuteczności algorytmu detekcji zasobów radiowych przed i po przeprowadzonych atakach.

Abstract: This paper presents the research results and analysis of the impact of poisoning label-flipping attacks on federated learning for spectrum sensing. The experiments have been executed for random and coordinated attacks for varying attackers-to-genuine-users ratios, different levels of aggressiveness, and time duration of attacks. The results have been obtained by comparing spectrum sensing machine learning model performance with and without attacks.

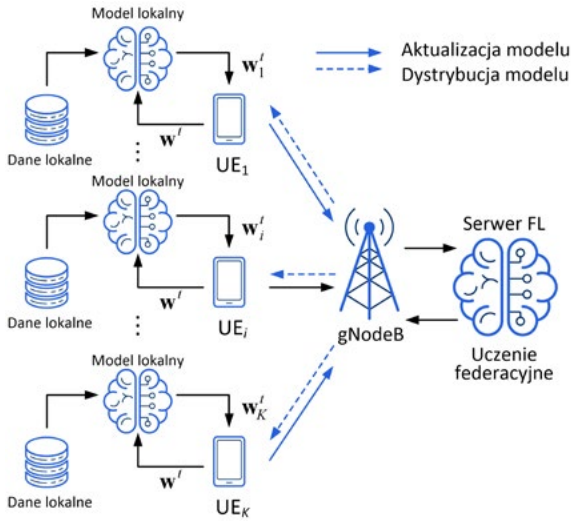
Słowa kluczowe: uczenie maszynowe, uczenie federacyjne, atak zatruwający, detekcja zasobów widmowych

Keywords: machine learning, federated learning, poisoning attack, spectrum sensing

1. WSTĘP

Wraz z coraz powszechniejszym użyciem metod uczenia maszynowego pojawia się problem pozyskiwania jak największej ilości danych treningowych oraz ich transferu w celu utworzenia dobrze wytrenowanych modeli. Transfer dużych ilości danych wymaga użycia odpowiednio dużo zasobów, takich jak energia, czas, a także zasoby radiowe w przypadku transmisji bezprzewodowej. Przesył danych wiąże się również z zagrożeniem ujawnienia danych wrażliwych użytkowników dokonujących transmisji. Jedną z odpowiedzi na te problemy jest algorytm uczenia federacyjnego (*ang. Federated Learning* -FL)

Dane zebrane przez użytkowników biorących udział w procesie FL charakteryzują się dużą różnorodnością, co odwzorowują lokalnie wygenerowane modele. Algorytm FL polega na tworzeniu w sposób iteracyjny globalnego modelu ML, dobrze uśredniającego panujące warunki w systemie, poprzez zebranie i zintegrowanie wielu modeli lokalnych wygenerowanych przez użytkowników [3]. W każdej iteracji globalny model jest rozsyłany do użytkowników biorących udział w procesie FL. Każdy z nich dokonuje zmian na otrzymanym modelu poprzez dalsze trenowanie go za pomocą lokalnie zebranych danych. Uzyskane w ten sposób modele są zbierane i agregowane w nową wersję modelu globalnego. Rysunek 1 przedstawia przykładową schemat wymiany informacji między użytkownikami, a serwerem FL, który jest odpowiedzialny za agregację modeli. Na przedstawionym schemacie systemu użytkownicy (UE), korzystają z zebranych danych, aby nauczyć się swoich lokalnych modeli. Wagi modeli z iteracji t oznaczone są jako w^t . Model globalny o wagach w^{t+1}, utworzony po uzyskaniu modeli o wagach w^t zostaje rozesłany do wszystkich UE na początku iteracji $t+1$ algorytmu FL.

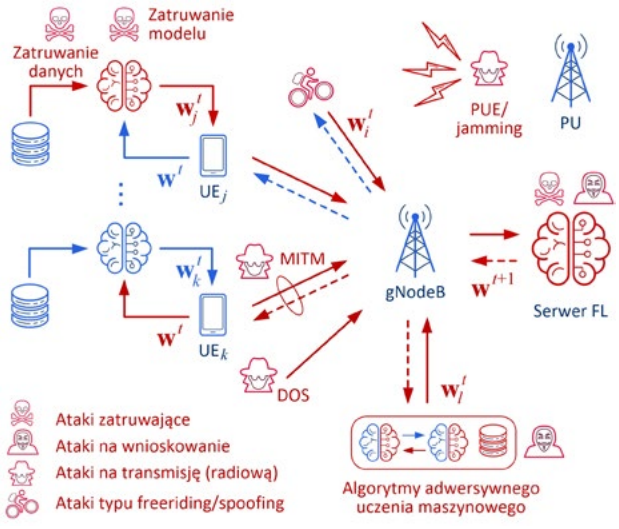


Rysunek 1: Schemat przedstawiający proces FL.

FL okazuje się szczególnie pomocne przy detekcji zajętości zasobów radiowych (*ang. spectrum sensing* - SS) [7, 6]. W przypadku FL użytego dla SS, użytkownicy (sensory) zbierają dane o zajętości kanału w różnych warunkach odbioru, tj. występują zaniki, shadowing, które mają wpływ na jakość odebranego sygnału. Częstość zmian w

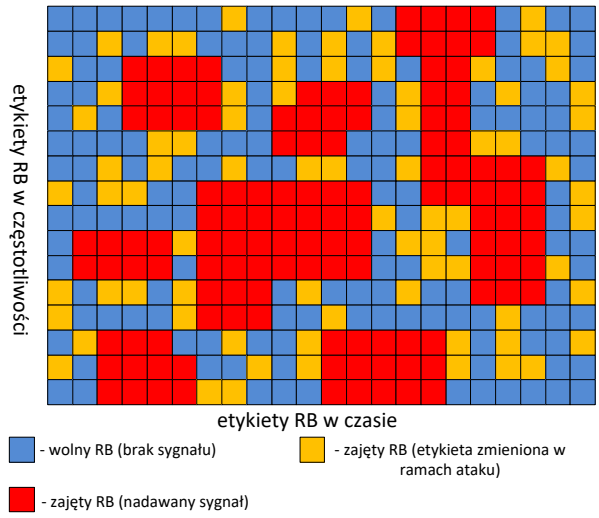
kanale również zależy od sensora, tj. od prędkości jego poruszania się względem źródła transmisji sygnału. Zastosowanie FL dla SS rozwiązuje problemy z indywidualnym SS przeprowadzonym za pomocą algorytmów ML. W indywidualnym SS wyżej wymienione ograniczenia w warunkach odbioru powodują niższą wiarygodność lokalnie wygenerowanego modelu ML, ponieważ użytkownik generujący lokalny model ML uczy się zależności w odbieranym sygnale, które są niejako zniekształcone przez panujące warunki odbioru. Intuicyjnym rozwiązaniem tego problemu jest zastosowanie scentralizowanego SS. To rozwiązanie jest jednak niepraktyczne ze względu na konieczność zebrania, przechowywania i przetwarzania naraz dużej ilości danych. Stąd FL dla SS poprzez swoje możliwości adaptacyjne do zmieniających się z czasem warunków propagacyjnych i przez rozłożenie zadania wygenerowania modeli ML na wielu użytkowników stanowi obiecującą odmianę kooperacyjnego SS.

W odróżnieniu od bezpośredniej wymiany danych treningowych, proces FL zapewnia wyższą ochronę danych wrażliwych, ponieważ w ramach tego procesu nie są wysyłane dane treningowe, ale wytrenowane modele. FL jest jednak podatny na ataki z zewnątrz ze względu na komunikację sensorów między sobą lub z serwerem FL, w którym odbywa się agregacja modeli. Możliwe ataki to m.in. ataki na komunikację między sensorami, a serwerem takie jak np. atak MiTM (*ang. Man-in-the-Middle attack*), czy jamming, ataki zatruwające na dane zbierane przez sensory, czyli zatruwanie danych (*ang. data poisoning*), ataki zatruwające model tworzony przez sensor (*model poisoning*) oraz ataki zatruwające na model globalny. Na rysunku 2 przedstawiono schemat komunikacji między sensorami, a serwerem FL oraz możliwe do przeprowadzenia ataki na różnych etapach algorytmu i na różnych elementach systemu. Więcej informacji dot. szczegółów możliwych ataków na FL dla SS można znaleźć w [10].

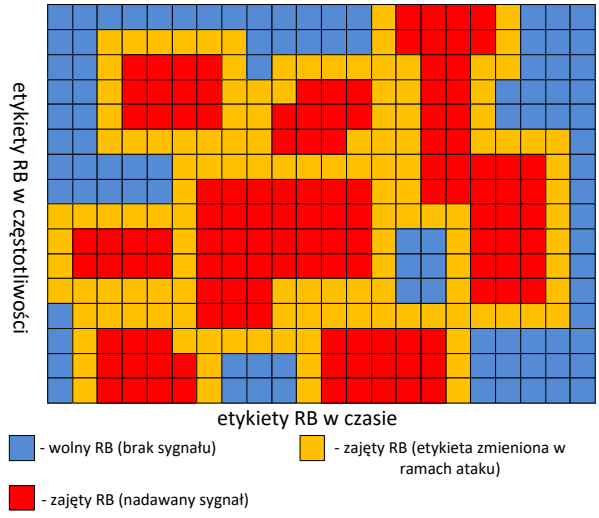


Rysunek 2: Rodzaje ataków na proces detekcji widma bazujący na FL.

W tym artykule skupiamy się na atakach polegających na celowej zmianie etykiet danych treningowych (ang. *label-flipping*). Ten typ ataków został opisany w pracach [4, 9, 2], które przede wszystkim dotyczą możliwości przeciwdziałania tym atakom. Wpływ ataków typu label-flipping został głębiej przeanalizowany w [8] oraz w [1], gdzie sprawdzono wpływ jaki ma procent zatrutych danych atakujących użytkowników na jakość modelu globalnego. Według naszej wiedzy nie ma artykułów, które dotyczyłyby wpływu jak ataki mają na skuteczność FL w zastosowaniu dla SS. Uważamy, że istotne jest, aby w badaniach nad atakami na algorytmy FL brać pod uwagę nie tylko metodę ataków, ale także motywację jaka stoi za wyborem działań atakujących. W przypadku FL dla SS atakujący przyjmą za cel nie tyle zakłócenie pracy działania algorytmu, co raczej próbę wprowadzenia w błąd innych użytkowników dokonujących SS. Celem takiego działania będzie stworzenie iluzji zajętego widma, tak aby inni użytkownicy rezygnowali z transmisji, co zapewni użytkownikom atakującym swobodny dostęp do zasobów radiowych. W tym artykule skupiamy się zatem na możliwych atakach przeprowadzonych na algorytm FL generujący modele dla SS.



(a) atak losowy



(b) atak przez enkapsulację (atak skoordynowany) pojedynczą warstwą RB

Rysunek 3: Porównanie etykiet danych treningowych w zależności od przyjętej metody ataku

Niniejszy artykuł uporządkowany jest w następujący sposób. W części 2 omówiono rozważane rodzaje ataków

zatruwających, których celem jest utrudnienie przeprowadzenia poprawnej detekcji wolnych zasobów. W części 3 przedstawiono scenariusz badań. W części 4 omówiony został eksperyment symulacyjny oraz wyniki będące rezultatem przeprowadzonych badań. Część 5 zawiera wnioski sformułowane dzięki analizie otrzymanych wyników.

2. SCENARIUSZ BADAŃ

Rozpatrywany jest system składający się z 25 sensorów i jednego serwera FL. Sensory dokonują detekcji energii sygnału stacji bazowej 5G. Wszystkie sensory poruszają się względem źródła sygnału z taką samą prędkością. Każdy z nich doświadcza przy odbiorze wpływu innego kanału, a więc innych zakłóceń odbieranego sygnału.

Rozpatrzono ataki zatruwające na dane zbierane lokalnie przez sensory. Taki atak może być również rozumiany jako atak na model lokalny tworzony przez dany sensor, ponieważ zakłamania w etykietach danych treningowych skutkują wygenerowaniem niepoprawnie działającego modelu. Równocześnie szkodliwy model lokalny wpływa w pewnym stopniu na model globalny.

W ramach rozważań sformułowano cel ataku na modele SS. Jest nim pozorne zwiększenie zajętości widma radiowego, co wiąże się z rezygnacją z transmisji użytkowników wtórnych. Skutkuje to zwiększeniem możliwości transmisyjnych dla użytkowników atakujących. Rozpatrzono ataki losowe i skoordynowane. W ramach ataków losowych sensory atakujące generują dane treningowe z etykietami pozmienianymi z "zasób wolny" na "zasób zajęty" dla losowo wybranych zasobów. Wszystkie sensory atakujące zmieniają jednak etykiety w taki sposób, aby osiągnąć zadany, nowy stopień zajętości widma wg nowych etykiet.

W atakach skoordynowanych wszystkie atakujące sensory zmieniają etykiety w taki sam sposób i dla tych samych zasobów w danych treningowych. Rodzaj oraz stopień agresywności ataku jest uzgodniony z góry. Zakładamy, że wszyscy użytkownicy (uczciwi oraz atakujący) mają dostęp do takich samych etykiet danych treningowych. Zatem uzgodniony najpierw rodzaj ataku stanowi dla atakujących skoordynowanych wystarczającą informację o tym, które etykiety mają zostać zmienione. Sygnałem podlegającym detekcji jest sygnał 5G, który w dużym stopniu charakteryzuje się doborem zasobów radiowych (*Resource Blocks - RBs*) występujących obok siebie zarówno w dziedzinie czasu jak i częstotliwości. Atakujący może wykorzystać tę wiedzę, aby w inteligentniejszy sposób zmodyfikować etykiety danych treningowych i tym samym przeprowadzić skuteczniejszy atak, niż zmieniając etykiety w sposób losowy. W ramach dalszych rozważań zaproponowano atak, nazwany przez autorów enkapsulacją, polegający po pierwsze na znalezieniu zgrupowań RBs oznaczonych poprawnie jako zajęte, a po drugie na zmianie etykiet przylegających do takich zgrupowań, tak aby niejako powiększyć obszar RBs uważanych w trakcie trenowania sieci za zajęty.

Na rysunku 3 przedstawiono przykładowy zestaw etykiet przyporządkowanych RBs w dziedzinie czasu i częstotliwości, a także wpływ obu metod ataku na etykiety użyte przez atakujące sensowy podczas generowania ich lokalnych modeli.

3. WYNIKI

W ramach badań, wygenerowano symulacyjnie sygnał 5G w paśmie 10MHz. Przyjęto model kanału typowy dla odbiornika poruszającego się z prędkością pieszego (*Extended Pedestrian A model*) [5]. Częstotliwość Dopplera została ustawiona na $2.5Hz$ dla wszystkich sensorów. Modele lokalne i globalne są generowane przy założeniu pewnego średniego SNR, wspólnego dla wszystkich sensorów.

Zestawienie parametrów eksperymentów wraz z parametrami rozważanych metod ataków przedstawiono w tabeli 1. Parametry wybrano w taki sposób, aby w sposób kompleksowy rozważyć różnorodne scenariusze ataków. Badania przeprowadzono w trakcie działania algorytmu FL, który został uruchomiony dla 15 iteracji indeksowanych od 0 do 14, dla różnych konfiguracji parametrów wejściowych. Jednym z nich jest stosunek liczby atakujących sensorów (K) do całkowitej liczby sensorów (N). Przetestowano również wpływ wyboru pierwszej iteracji algorytmu FL, od której rozpoczęły atak sensory atakujące. Pierwotny stopień zajętości widma danych treningowych i testowych wynosi 30%. W ramach ataku losowego zbadano zwiększenie stopnia zajętości widma do 50% i 70%. W ramach ataku skoordynowanego przetestowano dwa zakresy enkapsulacji za pomocą obudowania grup RBs pojedynczą warstwą RB, lub podwójną, które skutkują zwiększeniem pozornej zajętości widma odpowiednio do około 58% i 76%.

Wartościami służącymi do ewaluacji skuteczności przeprowadzonych ataków są dwa prawdopodobieństwa, które w sposób kompletny opisują jakość algorytmu SS. Są to prawdopodobieństwo detekcji P_{d} i prawdopodobieństwo fałszywego alarmu P_{fa}. P_{d}, dla naszych zastosowań, określa prawdopodobieństwo poprawnego zaklasyfikowania danego RB jako zajętego, pod warunkiem, że faktycznie jest on zajęty. Natomiast P_{fa} określa prawdopodobieństw z jakim dany RB został zaklasyfikowany jako zajęty, pomimo tego, że w rzeczywistości jest wolny. Dobrze działający algorytm SS maksymalizuje wartość P_{d}, przy jednoczesnej minimalizacji P_{fa}. Wysoka wartość P_{d} chroni sygnał podlegający SS, w razie podjęcia decyzji o transmisji w ramach zasobów uznanych za wolne. Niska wartość P_{fa} oznacza maksymalizację możliwości transmisji w wolnych zasobach.

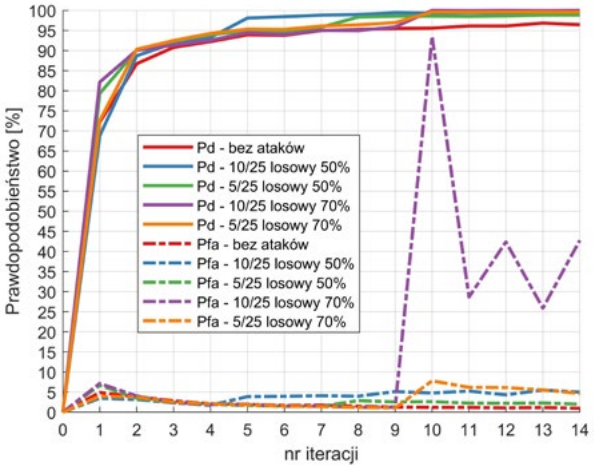
Tabela 1: Parametry i ich wartości wybrane do przeprowadzenia eksperymentów

Parametry symulacji	*Wybrane wartości*
SNR [dB]	0, 5, 10, 15, 20
maksymalny ind. iter. FL	14
Ataki - parametry	
Losowy	
- K/N	5/25, 10/25
- zajętość widma	50%, 70%
- iter. FL rozp. ataków	5, 8, 10
Skoordynowany	
- K/N	5/25, 10/25
- zakres enkapsulacji	1 RBs, 2 RBs
- iter. FL rozp. ataków	5, 8, 10

Rozważane algorytmy atakujące mają na celu pozorne zwiększenie częstotliwości klasyfikowania RBs za zaję-

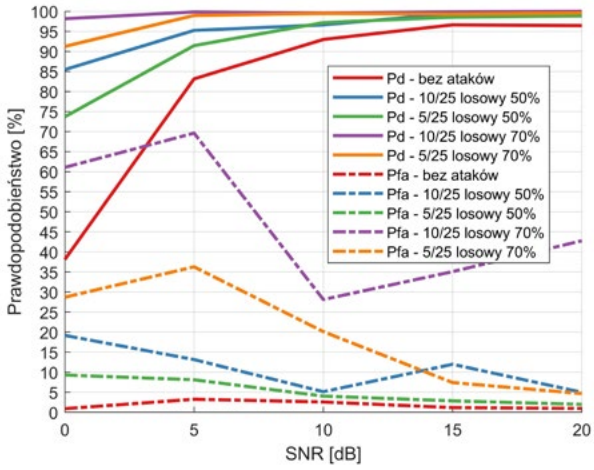
te. Algorytm atakujący powinien zatem podwyższać uzyskiwane wartości P_{fa} modelu globalnego.

Poniżej opisano wyniki symulacji przy założeniu parametrów z tabeli 1. Wyniki uzyskano poprzez wygenerowanie modeli globalnych dla wszystkich 14 iteracji, a następnie przetestowaniu ich na 15 zestawach danych wejściowych. Wyniki końcowe stanowią średnią uzyskanych wyników z tych 15 testów. Taki zestaw danych testowych okazał się wystarczający dla prawie wszystkich rozpatrywanych scenariuszy.



Rysunek 4: Prawdopodobieństwo detekcji i fałszywego alarmu dla FL poddanego atakom typu losowego dla SNR = 20 dB w zależności od iteracji FL

W pierwszej fazie badań przetestowano działanie ataku losowego i jego wpływu na P_{d} i P_{fa}. Na rysunku 4 przedstawiono wyniki dla SNR = 20 dB uzyskane po przeprowadzeniu ataku losowego w kolejnych iteracjach FL. Czerwonym kolorem zaznaczono wyniki P_{d} i P_{fa} dla systemu FL, w którym ataki nie występują. Pierwszy eksperyment przeprowadzono dla ataków dla $K/N = 5/25$, i zajętości widma 50%. Atak rozpoczyna się i trwa nieprzerwanie od iteracji nr 8. Atak losowy o takiej liczbie atakujących i zwiększeniu zajętości widma do 50% nie wpłynął w znaczący sposób na wyniki. Zauważalne jest niewielkie podniesienie wartości zarówno P_{d} i P_{fa} od iteracji nr 8. Można zauważyć, że wzrost P_{d} jest trochę większy niż wzrost P_{fa}. Jest to sytuacja korzystna dla zwiększenia ochrony transmisji sygnału, który podlega detekcji. Wynika to z tego, że podwyższenie wartości P_{d}, wiąże się z wyższymi wartościami P_{fa} i odwrotnie. W przypadku tego stosunkowo łagodnego ataku wpływ na wzmocnienie efektu fałszywego alarmu jest niewielki i nie przekroczył naturalnego zwiększenia poziomu detekcji. Podobne wyniki uzyskano dla ataku o większej liczbie atakujących tj. $K/N = 10/25$ i takiej samej zajętości widma 50%. Ten atak rozpoczęto od iteracji nr 5. Wartości P_{fa} są nieco wyższe niż dla poprzednio omówionego zestawu wyników, w związku z tym, że liczba atakujących jest większa w stosunku do całkowitej liczby sensorów. Można również zauważyć, że dla obu przypadków wybór iteracji rozpoczęcia ataków nie wpływa znacząco na wyniki.



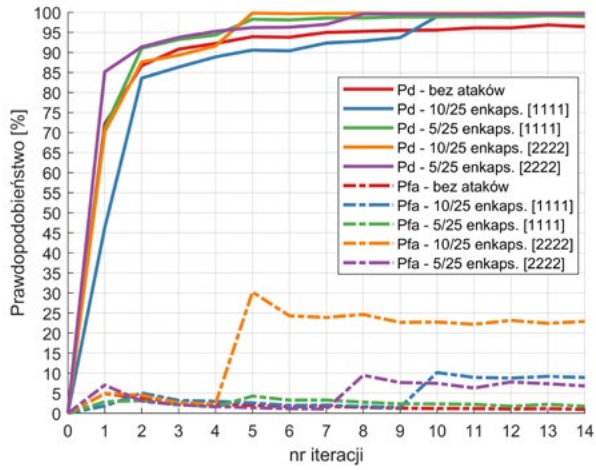
Rysunek 5: Prawdopodobieństwo detekcji i fałszywego alarmu dla FL poddanego atakom typu losowego dla ostatniej iteracji w zależności od SNR

Przeprowadzono również ataki dla takich samych wartości K/N, ale dla zwiększonej części zajętości widma tj. 70%. Ataki zarówno dla $K/N = 5/25$, jak i $K/N = 10/25$ rozpoczęto w iteracji nr 10. Dla $K/N = 5/25$ jest już zauważalny zwiększony wpływ na wartości P_{fa}, ale nie jest znaczący. Zupełnie inne wyniki uzyskano dla $K/N = 10/25$. W tym przypadku atak jest na tyle agresywny, że wartości P_{fa} wzrosły aż do prawie 95% w pierwszej iteracji ataku. W kolejnych iteracjach P_{fa} znacząco spada, ale aż do ostatniej iteracji nie osiąga stabilnego stanu, stąd duża zmienność wartości P_{d}. Atak w tym przypadku jest na tyle agresywny i zgodnie ze swoją losową naturą nieprzewidywalny, że wahania wyników są stosunkowo duże.

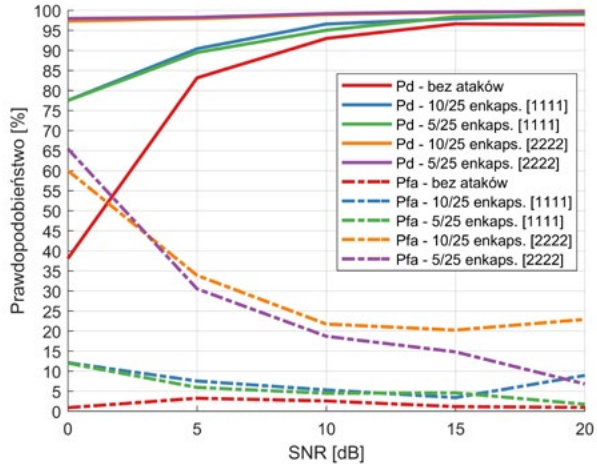
Na rysunku 5 porównano wyniki uzyskane dla ataków losowych w zależności od wartości SNR. Wszystkie parametry ataków są takie same jak dla wynikiów przedstawionych na rysunku 4, ale dotyczą ostatniej iteracji FL nr 14. Warto zauważyć, że im wartość SNR niższa, i im agresywniejszy atak, tj. więcej atakujących, lub większa część widma jest uznana za zajętą podczas trenowania atakujących modeli, tym wyższe warości zarówno P_{d} jak i P_{fa}. Niższa wartość SNR oznacza niższą dokładność algorytmu detekcji. W takich warunkach atak wpływa zarówno na poziom poprawnej detekcji jak i fałszywego alarmu. Podobnie jak dla wyników przedstawionych na rysunku 4 wartości prawdoposobieństw uzyskanych dla $K/N = 10/25$ i zajętości widma 70% charakteryzują się dużą zmiennością, wynikającąz losowej natury tych ataków. Aby nieco złagodzić ten efekt, potrzebne by było przeprowadzenie znacznie większej liczby symulacji.

Następnie przeprowadzono i zbadano wpływ ataków skoordynowanych. Rysunek 6 przedstawia wyniki w zależności od iteracji algorytmu FL. Podobnie jak w przypadku wyników dla ataków losowych, wyższy stopień agresywności zwiększa skuteczność ataku, tj. zwiększa wartość P_{fa}. Przeprowadzone ataki skoordynowane korzystają z wiedzy o transmitowanym sygnale, tak aby ich szkodliwe działanie było skuteczniejsze, czyli wyrządzało więcej szkód w modelu globalnym (podwyższało w większym stopniu P_{fa}), przy jednoczesnie mniejszej liczbie zmienionych etykiet i mniejszej liczbie atakujących UE. Ten efekt jest zauważalny dla wyników np. dla $K/N = 10/25$ i enkapsulacji o zakresie 1 RBs ([1111]), która odpowiada zwięk-

szeniem zajętości widma do 58%. Dla takiego ataku osiągnięto P_{fa} równe prawie 10%, podczas gdy dla ataków losowych o zajętością widma 50% uzyskano maksymalnie 7% P_{fa}. Analizując rysunek 6 i porównując go z rysunkiem 4 można dojść do wniosku, że po pierwsze, ataki skoordynowane są wrażliwsze na zmiany przyjmowanej zajętości widma i liczby K, a co za tym idzie bardziej precyzyjne i lepsze do użycia, w przypadku potrzeby osiągnięcia konkretnych wartości P_{fa}.



Rysunek 6: Prawdopodobieństwo detekcji i fałszywego alarmu dla FL poddanego atakom typu enkapsulującego dla SNR = 20 dB w zależności od iteracji FL



Rysunek 7: Prawdopodobieństwo detekcji i fałszywego alarmu dla FL poddanego atakom typu enkapsulującego dla ostatniej iteracji w zależności od SNR

Podobnie jak dla ataków losowych, również dla ataków skoordynowanych przeprowadzono badania dla różnych wartości SNR. Wyniki przedstawiono na rysunku 7. Warte zauważenia są wysokie wartości P_{fa} dla ataków o zakresie enkapsulacji 2 RBs z każdej strony ([2222]) dla niskich SNR. Uzyskane wartości P_{fa} są wyższe niż dla ataków losowych przy podobnym poziomie P_{d} szczególnie dla ataków z $K/N = 5/25$, dla których P_{fa} dla SNR=0 dB osiągnęło wartości ok 30% oraz 60% odpowiednio dla ataków losowych i skoordynowanych.

4. WNIOSKI

W niniejszym artykule przedstawiono porównanie ataków zatruwających typu label-flipping na dane treningowe w FL zastosowanym do tworzenia modeli SS. Ataki podzielono na losowe i skoordynowane i przeprowadzono analizę ich skuteczności, czyli uzyskanych wartości P_{d} i P_{fa}. Ataki miały na celu zwiększenie wartości P_{fa}, co przekłada się na częstsze, błędne uznanie widma za zajęte. Cel ten został osiągnięty. Ważne spostrzeżenia dotyczą także samych metod ataków, tzn. ataki skoordynowane charakteryzują się większą wrażliwością, czyli większymi zmianami w wynikach SS, w zależności od przyjętej zajętości widma i liczby użytkowników atakujących. Mniejsza liczba zmienionych etykiet wykorzystana w etapie trenowania sieci potencjalnie przekłada się na większe trudności w wykryciu ataku. To zagadnienie, czyli metody wykrywania przedstawionych w tym artykule ataków, a także metody usuwania, bądź minimalizowania ich skutków będziemy badać w przyszłych eksperymentach.

PODZIĘKOWANIA

Praca w części została sfinansowana przez Narodowe Centrum Badań i Rozwoju w ramach projektu 5GStar CYBERSECIDENT/487845/IV/NCBR/2021 pt. „Zaawansowane metody i techniki identyfikacji i przeciwdziałania cyberatakom na sieć dostępową 5G i aplikacje” oraz przez projekt 0312/SBAD/8164.

LITERATURA

[1] Aryal, K., Gupta, M., & Abdelsalam, M. (2022). Analysis of Label-Flip Poisoning Attack on Machine Learning Based Malware Detector. In 2022 IEEE International Conference on Big Data (Big Data) (pp. 4236-4245). IEEE.

[2] Li, D., Wong, W. E., Wang, W., Yao, Y., & Chau, M. (2021, August). Detection and mitigation of label-flipping attacks in federated learning systems with KPCA and K-means. In 2021 8th International Conference on Dependable Systems and Their Applications (DSA) (pp. 551-559). IEEE.

[3] Chen, M., Gündüz, D., Huang, K., Saad, W., Bennis, M., Feljan, A. V., & Poor, H. V. (2021). Distributed learning in wireless networks: Recent progress and future challenges. IEEE Journal on Selected Areas in Communications, 39(12), 3579-3605.

[4] Richards, C., Khemani, S., & Li, F. (2022, October). Evaluation of Various Defense Techniques Against Targeted Poisoning Attacks in Federated Learning. In 2022 IEEE 19th International Conference on Mobile Ad Hoc and Smart Systems (MASS) (pp. 693-698). IEEE.

[5] Song, L. & Shen, J. (2010). Evolved Cellular Network Planning and Optimization for UMTS and LTE. CRC Press: Boca Raton, FL, USA, 56–58.

[6] Chen, Z., Xu, Y. Q., Wang, H., & Guo, D. (2021). Federated learning-based cooperative spectrum sensing in cognitive radio. IEEE Communications Letters, 26(2), 330-334.

[7] Wasilewska, M., Bogucka, H., & Kliks, A. (2021). Federated learning for 5G radio spectrum sensing. Sensors, 22(1), 198.

[8] Uprety, A., & Rawat, D. B. (2021). Mitigating poisoning attack in federated learning. In 2021 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 01-07). IEEE.

[9] Liu, X., Li, H., Xu, G., Chen, Z., Huang, X., & Lu, R. (2021). Privacy-enhanced federated learning against poisoning adversaries. IEEE Transactions on Information Forensics and Security, 16, 4574-4588.

[10] Wasilewska, M., Bogucka, H., & Poor, H. V. (2023). Secure Federated Learning for Cognitive Radio Sensing. IEEE Communications Magazine, 61(3), 68-73.

AI-ASSISTED DIMENSIONING OF 5G NETWORK SLICES – REVIEW AND PERSPECTIVES

Dominik Dulas[1,2]; Krzysztof Walkowiak[2]

[1] Nokia Solutions and Networks, dominik.dulas@nokia.com
[2] Wroclaw University of Science and Technology, Wroclaw, Poland, krzysztof.walkowiak@pwr.edu.pl

DOI: 10.15199/59.2023.4.43

Streszczenie: **Zastosowanie plastrowania sieci 5G umożliwi stworzenie zestawu nowych usług celem automatyzacji przedsiębiorstw oraz ułatwienia naszego życia codziennego. W tym artykule, przedstawiamy przegląd wyzwań, analizujemy obecny stan wiedzy oraz wskazujemy perspektywy w zakresie wymiarowania plastrów sieci 5G wspomaganego metodami sztucznej inteligencji. Pomimo wielu badań, w szczególności w zakresie predykcji ruchu i zarządzania plastrami wspomaganych metodami AI, w obecnych badaniach naukowych nie znajdujemy prac przedstawiających całościowe rozwiązanie do wymiarowania plastrów sieci 5G w chmurze.**

Abstract**: 5G network slicing applications enable a set of new services that will automate industries and enhance our daily life. In this paper, we provide a review of challenges, investigation of the state of the art and show perspectives for future work in the area of AI-assisted dimensioning of 5G network slicing. Although there is a lot of research done, especially in the aspects of traffic predictions and network slicing management with AI assistance, in the current international publications there is no holistic solution for dimensioning of 5G radio cloud with network slicing.**

Słowa kluczowe**: bezprzewodowe sieci 5G, wymiarowanie, plastrowanie sieci, QoS.**

Keywords**: 5G wireless networks, Cloud RAN, Dimensioning, Network Slicing, QoS.**

1. INTRODUCTION

Enhanced mobile broadband (eMBB) is the evolution of next generation mobile networks, but 5G is much more than higher speeds and greater capacity. 5G architecture with lower delays and higher availability creates the possibility to establish ultra-reliable, low-latency communications (URLLC) and massive machine communications (MMC) for mission-critical applications [2][14]. These applications enable a set of new services that will automate industries and enhance our daily life. However, these applications implicate a new, broad range of requirements on 5G network that will need to be provided simultaneously.

The solution to manage a new range of needs is network slicing, which is essentially a network virtualization technique [8]. Each slice is a logical network on the top of a physical network, that contains all necessary resources to serve connected applications and/or users with specific requirements. All physical network resources are assigned to the slices in a shared and/or dedicated manner. Dynamic slice resources assignment will improve the efficiency of the network utilization and enable scaling. However, network slicing does introduce additional management challenges and Nokia Bell Labs estimates that the added complexity of performing network slicing manually will increase the total cost of ownership by 30 percent versus that of traditional network [7].

Communication service providers (CSP's) responsibility is to properly plan, deploy and maintain networks consisting of hundreds to thousands of radio base stations. Network slicing with a new logical layer of competing "sub-networks" adds an additional layer of complexity to these processes.

This paper provides a comprehensive review of the network capacity planning challenges for CSPs' introduced by network slicing, followed up by recent advances and future trends how AI can be used to overcome them.

The paper is organized as follows. Section 2 outlines the dimensioning process and its current challenges. Section 3 describes how AI may help and how it is used in recent works. Section 4 concludes this study.

2. 5G DIMENSIONING CHALLENGES

2.1. One fit all process

One of the elements of the Radio Access Network (RAN) planning process is the capacity dimensioning [14]. The objective is to estimate the number of resources required to provide a service for specific traffic demand and with proper quality of service (*QoS*). In general, it consists of the following stages presented on Fig. 1. They can be performed one-by-one or in parallel.

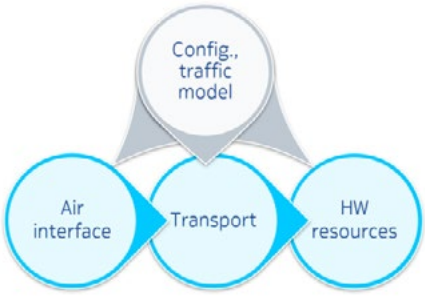


Fig. 1 Dimensioning process flow.

At start, an air interface dimensioning needs to be performed in order to provide rough estimation of the site volumes when deploying a radio network. Next, transport

dimensioning aims to estimate required capacities on access network interfaces[9]. Finally, hardware (HW) resources dimensioning is done to obtain required number of physical resources, counted in modules (e.g. radio, system) for classical bare metal deployments and in central processing units (CPUs) for cloud deployments.

All of these stages require definition of common inputs, like network configuration (e.g. cell bandwidth, MIMO mode, functional split) and traffic model (e.g. number of subscribers, data volume), which depend on the CSP requirements (coming from technology vision).

Dimensioning process is done by vendor proprietary tooling, since all the specific equipment limitations are internal vendor know-how. Currently, it often uses spreadsheets and lacks a more sophisticated approach [18]. It also requires expert know-how from planning engineers to perform traffic forecasting and performance prediction. Manual or so called "spreadsheet" planning makes the dimensioning of whole network with heterogenous configurations and differentiated traffic models and its services harder or even unfeasible. This approach is ineffective and leads to "one rule fits all" principle that assumes all cells to be equal. Finally, manual planning is also prone to human error, so the result might be inaccurate.

It is worth to distinct two cases for the dimensioning. A long term case is needed once a CSP is planning CAPEX investments (usually for 1-3 years) and will select vendor(s) that will provide equipment for network deployment. In such a case, tooling should estimate enough resource planning for future needs. A short term case is needed for a daily operation of network with slicing, because all the slices compete for the limited number of the resources (estimated in the long term case) according to momentary needs. Both cases result in the same outputs (site capacities) but due to different time scale may require different approaches.

2.2. A priori defined traffic model

Dimensioning process requires input data, describing the conditions in which network shall operate and requirements from services it shall provide. Traffic model (TM) is its numerical description. Traffic data is forecasted to define TM for some time in the future when network is planned to be operational [10].

CSPs provide number of subscribers for next few years that they plan to have in their network. Service mix and service demand per subscriber is also provided, but usually assumed from some generic, averaged, statistic-based model, usually also linearly extrapolated in time. This may be fine for a long term planning, but will not be precise for site specific requirements and also for short term estimations of time-variant wireless networks (especially in network slices with high mobility) [2].

Although having very precise dimensioning tooling, the result will be as good as accurate TM is provided. Therefore, it is most important to extend the dimensioning process with traffic forecasting methods based on real network data or design the network without the need of TM [17].

2.3. Linear models

Mobile networks dimensioning models are created based on network simulations while the product is in the development phase and enhanced with test results when the product is ready. Both simulations and tests require time and resources and in practice can be done only for a subset of all possible scenarios (meaning the combination of network configuration, feature activation and traffic mix). As a result, standard linear or queueing models are created that match simulated and tested scenarios [4].

Any other (un-simulated or un-tested) scenario, especially with complex configuration and combination of features, will decrease the accuracy of the dimensioning model. Network slicing adds another dimension to the complexity of the process as the traffic mix and related service requirements are multiplicated for the specific scenario.

Impacts of individual product features are added to the model as additional linear model or a coefficient. Not all features are independent and often they impact each other performance. Therefore, so called "feature bundles" shall be simulated or tested which increases the efforts to create dimensioning model. Sum of coefficients or feature specific linear models may give inaccurate results due the effect of multicollinearity [18].

Applying existing standard linear models in the dimensioning process yields accurate results for specific scenarios. However, with the ever-increasing number of configurations and features, sum of scenario specific linear models will result in decreased process reliability. These observations triggers the work in a range of non-linear modeling, including AI-oriented techniques. Furthermore, applying existing iterative model-based dimensioning methods can be unsuitable since the computational complexity greatly increases with the network scale [2].

2.4. Cloud

Network Functions Virtualization (NFV) main advantage is that it enables the operator to scale resources up and down faster than in the traditional Physical Network Functions (PNF), where the hardware had to be procured, commissioned, and connected to be made available for deployment of a new network application. Virtual Network Functions (VNF) are then deployed as SW applications on top of a Network Functions Virtualization Infrastructure (NFVI) to provide their telco services at the operator's premises [2].

Cloud RAN enables a split of Base Transceiver Station (BTS) into Radio, Distributed and Centralized Units (RU, DU and CU) that can be geographically separated (Fig. 2), e.g. to enable resource pooling. Whereas in classical deployments, fronthaul (FH) and midhaul (MH) links are usually realized over physical point-to-point links as all elements are collocated. That's why CSP's need to consider FH and MH link capacity dimensioning on top of Backhaul (BH).

The last step of the dimensioning process (Fig. 1) is straightforward in classical RAN deployments where software (SW) is tight to HW. Because of that, capacities of SW+HW are known in all dimensioning phases. However, for cloud dimensioning additional step needs to be

done, where virtual resources are mapped to HW resources.

2.5. Access transport aggregation

Traffic transmitted via the radio interface is sent over the access transport network towards and from core network elements creating aggregation points (Fig. 2). Transport links must provide sufficient capacity with *QoS* to enable the required radio performance. Therefore, part of the RAN dimensioning process is to estimate the capacity of individual BTS interfaces and aggregation points in access domain. It is worth mentioning that due to the burstiness of the packet traffic it is a place of potential cost saving thanks to the statistical Multiplexing Gain [9].

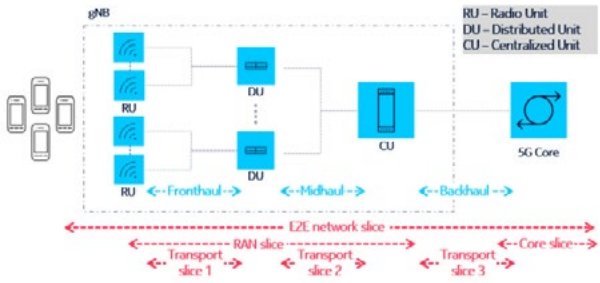


Fig. 2 5G cloudified gNB with transport interfaces and network slices.

Again, to understand the *QoS* requirement (temporal values of throughputs) of the radio interface and traffic profile patterns for specific scenarios, the network system level simulator or network level tests have to be done in traditional dimensioning approach (as described in section 2.3). They provide the basis for linear model creation per scenario assuming homogeneous network configuration and traffic requirements. This obviously is a place for potential improvement.

Additionally, due to the cloudification and split of BTS functions, transport (especially FH) is impacted depending on the functional split between RU and DU [15]. At the current phase of 5G deployment, FH link capacity depends on the split in a static way. At the next stage, with the evolution of FH functionalities and the need to reduce its capacity, it will be dependent of the traffic it transmits.

Transport planning is an integral part of overall Radio Access Network (RAN) planning as cost optimization of transport network may lead to cost reduction of the whole mobile network deployment. However, transport links must provide sufficient capacity and Quality of Service (*QoS*) to enable required radio performance (which is usually part of CSP service commitments and marketing). Therefore, the cost reduction on transport network is demanded but not always for the price of radio interface performance.

2.6. Network slicing

Introduction of network slicing adds another layer of complexity to the aforementioned challenges. As each slice may have different requirements (TM, *QoS*) and the network resources are shared, all those requirements have to be taken into account in dimensioning.

Network slice (end to end – E2E) is divided into RAN slice, core slice and transport slices [15] (see Fig. 2). Although slice requirements are defined E2E, each slice part is dimensioned separately, because of specifics of underlying resources (radio, transport, core). Therefore, each slice part will require dedicated model.

3. AI PERSPECTIVES

Dimensioning challenges described in previous chapter can be overcome thanks to AI techniques and below we show examples in which cases and how it is done at the current state of the art.

3.1. Process based on heterogenous data

Machine learning (ML) algorithms enable modelling of each cell or cluster of similar (from configurational and/or performance perspectives) cells individually, based on its unique characteristics [2]. Multivariable modeling techniques consider multiple factors to predict traffic growth per cell. This approach enables precise, heterogenous planning of network resources with local requirements and eliminates the “one rule fits all” method.

ML algorithms need data for training. Network data coming from simulations or tests (used in majority of research) shows only a small fraction of possible configurations and traffic loads. Therefore, it would be best to do the training on data from live commercial networks. In [1], authors propose a framework based on continuous deployment in commercial network to create ML based models.

3.2. Forecasted traffic model

Historical data can be the input to service demand prediction via AI techniques. Recent studies show that AI-based methods, such as deep neural network (DNN), long short-term memory (LSTM) and recurrent neural networks (RNN), are capable of accurately forecasting service-specific traffic load [2].

For the short term prediction of PRB (Physical Resource Block) resource usage by a slice, a realistic time series prediction model based on X-LSTM and validated in custom LTE testbed have been proposed in [12]. Service level traffic demands prediction within individual network slices using a deep learning framework is proposed in [5]. Thanks to the reduction of resource overprovisioning and service request violations it allows substantial operating costs reduction. Service traffic prediction is also described in [13], here with unsupervised learning model. In [6], spatial and temporal correlation of data traffic arrivals has been exploited to jointly predict the traffic arrivals at all base stations with a DNN.

Enlisted existing studies show the potential of AI-assisted forecasting methods to accurately predict traffic demands on slice level. Such approach can improve the accuracy of the traffic model used in the dimensioning process.

Authors current research aims at design of a multi-dimensional model based on the time series methodology that could be used for the 5G QoS/slice level traffic forecasting. We assess the accuracy of VARMA, Prophet and neural networks models for this task.

3.3. Slicing management

Managing a number of network slices creates significant cost that can be reduced by AI. It can be used as a tool in all phases (preparation, planning and operation) [4]. It has the potential of handling complicated decision-making problems in a dynamic network environment e.g. for transmission power allocation in cellular networks and resource allocation in network slices [8]. As an example, the game theory is used in a framework for planning RAN slices with Guaranteed Bit Rate in [11]. Supervised DNN is proposed for spectrum allocation, aiming to minimize costs, maximize radio resource utilization, and guarantee desired service level agreements in [16].

4. CONCLUSIONS

Assistance of AI in radio capacity dimensioning is becoming a requirement due many reasons and challenges mentioned in this paper. 2G, 3G, 4G dimensioning methods were based on simulation or test data and linear models with feature coefficients. With 5G and especially network slicing, system and processes complexity are highly increasing with the introduction of cloud and network slicing. Additionally, for short term dimensioning there is a need for effective solutions with fast response and great accuracy to dynamically deploy network slices.

Solutions based on network data (esp. commercial) may only cope with such requirements. Intelligent methods, thanks to the aspects of training process, guarantee higher scalability (small, large networks) and elasticity (transfer of trained model to different network location, other services). Once trained, they will provide fast response to queries. Finally, it can be re-trained with the evolution of network, traffic model, introduction of new slices, etc.

Although there is a lot of research done, especially in the aspects of traffic predictions and network slicing management with AI assistance, in the current international publications there is no holistic solution for dimensioning of 5G radio cloud with network slicing. We envision that future work will extend the usage of commercial data in the training process. Additionally, methods for traffic forecasting will be joined with methods for dimensioning (learning and predicting system capacity). Finally, there will be new area of research related to explainable AI, to make the results of these methods understandable for CSPs.

ACKNOWLEDGMENT

The research leading to these results has received partial funding from the Polish Ministry of Education and Science via the Industrial PhD program.

REFERENCES

[1] Panek M. et. al. 2023. „5G/5G+ network management employing AI-based continuous deployment". Applied Soft Computing, Volume 134.
[2] Barakabitze A. A. et. al. 2020. "5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges". Computer Networks, Volume 167.
[3] Shen X. et al. 2020. "AI-Assisted Network-Slicing Based Next-Generation Wireless Networks". IEEE Open Journal of Vehicular Technology, vol. 1.
[4] W. Wu et al.. 2022. "AI-Native Network Slicing for 6G Networks". IEEE Wirel. Commun., vol. 29.
[5] Bega D. et. al. 2019. "DeepCog: Cognitive Network Management in Sliced 5G Networks with Deep Learning". IEEE INFOCOM 2019, Paris, France.
[6] J. Ye and Y. -J. A. Zhang. 2020. "DRAG: Deep Reinforcement Learning Based Base Station Activation in Heterogeneous Networks". IEEE Trans. Mob. Comput., vol. 19.
[7] Narayan Raman et al. 2018. "Future X Network Cost Economics - A network operator's TCO journey through virtualization, automation, and network slicing". Bell Labs Consulting.
[8] X. Shen et. al. 2022. "Holistic Network Virtualization and Pervasive Network Intelligence for 6G". IEEE Communications Surveys & Tutorials, vol. 24.
[9] Dulas D. et. al. 2022. "Method of 5G TDD Midhaul Multiplexing Gain Estimation based on System-Level Traffic Measurements". SoftCOM, Croatia.
[10] Kirmaz A. et. al. 2020. "Mobile Network Traffic Forecasting Using Artificial Neural Networks". MASCOTS 2020.
[11] Adamuz-Hinojosa O. et. al. 2023. "Potential-Game-Based 5G RAN Slice Planning for GBR Services". IEEE Access, vol. 11.
[12] Gutterman C. et al. 2019. "RAN Resource Usage Prediction for a 5G Slice Broker". Mobihoc '19. New York, USA.
[13] Sciancalepore V. et. al. 2019. "RL-NSB: Reinforcement Learning-Based 5G Network Slice Broker" IEEE/ACM Transactions on Networking, vol. 27.
[14] Khan M. U. et. al. 2020. "Service-Based Network Dimensioning for 5G Networks Assisted by Real Data,". IEEE Access, vol. 8.
[15] Touati H. et. al. 2020. "Split analysis and fronthaul dimensioning in 5G C-RAN to guarantee ultra low latency". IEEE CCNC, USA.
[16] Abozariba R. et. al. 2020. "Uncertainty-aware RAN Slicing via Machine Learning Predictions in Next-Generation Networks". VTC2020-Fall, Canada.
[17] Bektas C. et. al. 2022. "The Cost of Uncertainty: Impact of Overprovisioning on the Dimensioning of Machine Learning-based Network Slicing". FNWF, Canada.
[18] https://www.ericsson.com/en/blog/2022/3/next-gen-ai-driven-network-dimensioning-solution

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

Wykorzystanie metod nienadzorowanego uczenia maszynowego do optymalizacji kolejki testów regresyjnych w stacjach bazowych 5G

An unsupervised machine learning approach for regression testing queue optimisation in 5G base stations

Sebastian Zarębski [1,2]; Marcin Jokiej [3]; Marcin Szczukiewicz [4]; Krzysztof Rusek [5]; Piotr Chołda [6]

[1] NOKIA, Kraków, sebastian.zarebski@nokia.com
[2] AGH University of Krakow, Kraków, zarebski@agh.edu.pl
[3] NOKIA, Wrocław, marcin.jokiej@nokia.com
[4] NOKIA, Wrocław, marcin.szczukiewicz@nokia.com
[5] AGH University of Krakow, Kraków, krzysztof.rusek@agh.edu.pl
[6] AGH University of Krakow, Kraków, piotr.cholda@agh.edu.pl

DOI: 10.15199/59.2023.4.44

Streszczenie: W organizacji działającej w oparciu o zwinne podejście do rozwoju oprogramowania, gdzie cykle życia oprogramowania są bardzo krótkie (np. dwa tygodnie), zmiany w oprogramowaniu są bardzo częste. Zazwyczaj zasoby są ograniczone — prąd jest drogi, linie testowe są stale zajęte, a części sprzętu muszą być zarezerwowane tylko do testów regresyjnych. W tej perspektywie testy regresyjne mogą wprowadzić wiele niepotrzebnych kosztów całkowitych. Porównując metody statystyczne oraz nienadzorowanego uczenia maszynowego odkryliśmy, że dzięki jednolitej naturze zmian w kodzie, można łatwo osiągnąć 90% dokładności przewidywania błędów przy jednoczesnym zmniejszeniu pierwotnej kolejki testów o 25%.

Abstract: In organization with applied agile software development, where software life cycles are very short (i.e. two weeks), changes to the software are very frequent. Usually resources are scarce – power is expensive, test lines are constantly occupied, and hardware parts must be booked only for regression testing. In this perspective, regression testing might introduce a lot of unnecessary overhead. By comparing statistical methods and related to unsupervised machine learning methods, we discovered that due to a uniform nature of code changes, one can easily achieve 90% of bug prediction accuracy while reducing the original testing queue by 25%.

Słowa kluczowe: testowanie oprogramowania, wybór przypadków testowych, ciągła integracja, testy regresyjne

Keywords: software testing, test case selection, continuous integration, regression testing

1. INTRODUCTION

Continuous integration is focused on automatically building and testing new code. This process must be accompanied by continuous delivery, which automates the entire software release process. NOKIA has its fully automated, resulting in a software development process that continues continuously throughout the week. As the new change is introduced into the code, regression testing is applied to verify whether the new implementation did not introduce a new bug. Even if the immediate goal of this kind of testing is not to eliminate all bugs (which can only be achieved through subsequent phases of the process, such as performance testing), the expectation is to maintain the consistency and sustainability of the software's core functions. It is therefore necessary to prepare, maintain, and continuously adjust the pool of tests that will satisfy these expectations. In extreme cases, as during the evolution of technology from the 4G to 5G standard, both technologies must be supported for a period of time. This leads to a significant increase in testing workload and test line occupancy.

We have information on not only the test results but also the exact daily software changes that were associated with those results. Therefore, we know not only what went wrong but also roughly why something failed to work. Therefore, the purpose of this study was to verify whether historical data obtained during the past year could be used as a source of information for intelligent management of the regression test pool. To verify this idea, we compared two approaches, the first purely statistical, using only the history of test results, and the second, in which we used unsupervised machine learning algorithms to find similarities in changes between two software versions and correlate them with test results.

We give the following structure to this paper. In Section 2, we present the background of the problem and the corresponding literature that inspired our idea. In Section 3, we describe the source of the problem and the implications associated with it. In Section 4, we describe the methods we used. In Section 5, we present our results and findings and discuss the findings in relation to possible future work. In Section 6, we conclude our research.

2. BACKGROUND AND LITERATURE REVIEW

The problem of optimizing the test queue associated with the software development life cycle has been addressed by researchers for many years. One type of test that, by its nature of repeatability, seems perfectly suited to automated management is regression testing.

Regression testing can be achieved through different approaches. We consider the queueing and prioritization of such tests [1]. Another option for managing a queue of

regression tests is to prioritize them according to a selected rule to mark their importance. One suggestion is to try to identify not specific classes of errors, but rather their cliques, based on their relation on the distance function between them [2].

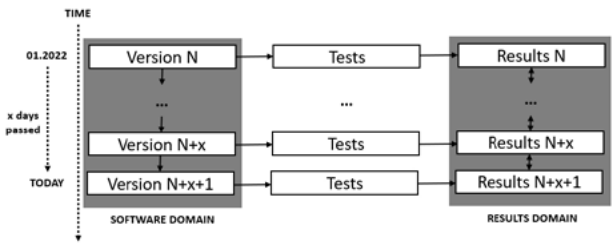
Several researchers applied different methods of test case selection or prioritization (TSP). These methods are supervised and unsupervised. For the first one, most of the papers define a TSP as a ranking problem. Several different methods were used to provide priority scores or probability of failure for ranking test cases, the most noticeable of which are logistic regression [3][4], SVM [5], XGBoost [6], RNN [7] and a Bayesian network [8]. Most of the reported models that use machine learning techniques are restricted by the classical ML problems, such as setting hyperparameters like batch size or number of neurons in layers. Others suffer from the problem of need to incrementally adjust model (continuous learning and retraining) and must be recreated from scratch in each training cycle. In the setting of regression testing, this may include daily recreation of the model, which may be cost- and time-consuming.

On the other hand, unsupervised approaches are also considered, in which clustering is widely adopted option. Clustering is widely used in general to identify and group similar data points into clusters, and these can be test cases or different versions of a software. As training data, the family of features called a 'code complexity' is commonly used. Due to simplification of computations, researchers tend to use Lines of Code (LoC) as the primary feature. As for the clustering methods, there are several approaches that the use of which brings measurable benefits. Noticeably the K-means algorithm is used [9], hierarchical clustering [10], or expectation maximization (EM) and density-based spatial clustering of applications with noise (DBSCAN) [11].

Beside the problem of finding an optimal number of clusters, there is a need to make a decision on how the distance between instances is calculated. Most studies rely on the use of the Euclidean distance [9] and some use the Hamming distance [10].

3. PROBLEM STATEMENT

Looking at the problem from a data perspective, one can see that it involves two domains. We refer to the first domain of **software changes** and the second, with regard to **testing results**. A specific relationship occurs in both areas, that some errors may be due to previous results (e.g., inaccurate reporting) or the same error may be so difficult that it is not fixed by several subsequent versions of the software. This relationship is shown in Figure 1.



Rysunek 1: Data domains related to the problem

Each software version is tested by a significant number of regression tests (several hundred to more than a thousand, depending on the hardware version or the complexity of the product). To limit the scope of the problem, we do not deal with the content of the tests, so we do not expand the subject area to natural language processing. Therefore, we treat each **test** as an single abstract object with two possible states: **failed** or **passed**. If there is no information on an outcome, we assume that the test passed.

Looking at two successive versions of the software, we can determine their difference, which we will call **delta** in the following parts of the article. A **software delta** is a numerical vector that describes the number of changes between components (let us assume classes in the code, which perform specific functions). Assuming that the history of changes is logged in version control systems (e.g. Git or SVN), we have an accurate count on the number of added and deleted lines of code in each of these components. In this way, we can observe the dynamics of changes in the software and tie it to their testing results.

Considering the above, we note that the problem is described by two matrices of known and constant shapes. The first one, related to software deltas, is the numerical matrix that describes a total number of changes per component (columns) in each consecutive software version (rows). We conclude that the problem is of type **binary classification**, in which the software delta matrix is the training data and the results of the tests describe the answers.

As the problem directly impacts the business side of the company, it was necessary to define the expectations and related quantifiable metrics. The first is the metric introduced solely for this study, called an optimization score. The value that describes the size of the subset of the original queue of tests that could be performed using only prediction of proposed methods. It is described by an equation $P = 1 - \frac{t_p}{t_s}$, in which t_p is the number of tests predicted to fail and t_n is the number of all tests in the scope. Other metrics are widely known that describe the outcome of the classification: F1 score, AUC-ROC, sensitivity, specificity and two accuracy scores (for all test failures and for actual bugs found — not all problems reported by tests are related to an actual bug in the code, e.g. problems with connectivity timeouts due to temporary connections lags).

Two different data sets are used. First one is related to testing history and has a shape of tens of thousands results (two thousand unique tests), each described by a test name, result, start time and duration and (if applicable) the bug's ticket id. The second one describes the changes in the software. It has about five hundred rows (per each day's software version under test) and about eight thousands columns describing changes in it to the previous version (sum of lines of code changed).

4. METHODS

For the sake of making the possible solution as simple and affordable as possible, we decided to focus on two approaches. First, statistical, aims to verify whether any machine learning or any other complicated approach is actually feasible. The second one focuses on two-step process — to find similar software versions with clustering method (unsupervised learning) and then pick tests that ended with failure for those software versions. One simple approach is considered – **baseline**, called internally a 'hi-

storical' approach: it uses historical data of failed results from a rolling time window from selected period of time.

For each of the **clustering method**, we use data from both domains, that is, from the software and the results domains. The first problem to consider is the use of a proper distance metric. There are two common and reasonable approaches, and either one is able to outperform the other empirically. The Euclidean distances work with the assumption that data is isotropic Gaussian (treating each feature equally), while the Mahalanobis distance is used when measuring the correlation between features and at the same time relaxes the mentioned assumption, assuming an anisotropic Gaussian distribution instead. Thus, if there exists a correlation between features, the Mahalanobis distance seems to be more plausible. The major drawback of its is, however, that it requires the inversion of the covariance matrix during computation, which can be computationally restrictive. Additionally, the covariance matrix must be positive and symmetric. It is worth noting that the Gaussian mixture model is using this distance.

To find similarities between two software versions, we decided to use the following methods: **hierarchical clustering** (agglomerative), ***K*-Means**, **Gaussian Mixture Model** and **DBSCAN**.

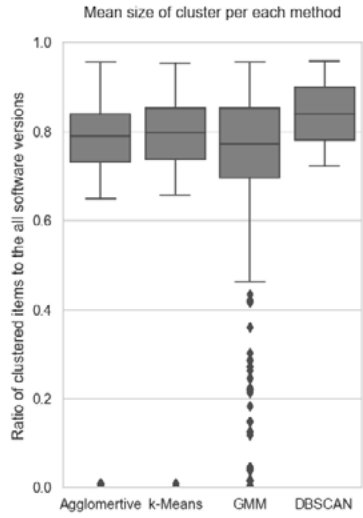
The first three algorithms are defined by the parameter of the target number of clusters. Optimal number of clusters is unknown; therefore, in order to test the effectiveness of each method, it was attempted to collect results for a range of clusters from 2 to the square root of the number of software versions under testing. Based on this, the mean and standard deviation values for each software version tested were determined according to the number of clusters and then in relation to the following clustering metrics: **Davis–Bouldin (DB) index** and the **Silhouette score**.

5. EMPIRICAL FINDINGS

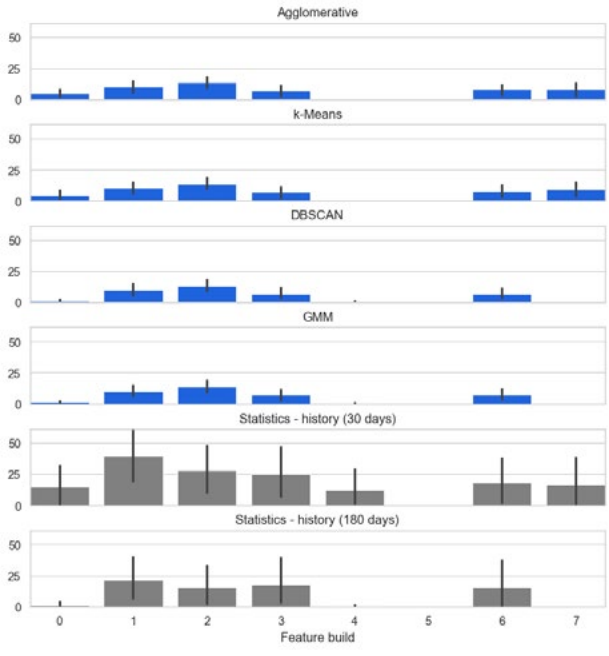
It was found that both clustering and statistical methods have a number of significant advantages and disadvantages, which do not allow to choose unequivocally the best method. The results obtained allowed us to conclude that any of the solutions used achieves a level of accuracy of at least 90% in terms of selecting potentially problematic tests. A more difficult task turns out to be finding the remaining 10%, and in this area each method made slightly different mistakes; therefore, the next part of this section focuses on the wrong classification in the context of possible optimization of the queue of regression tests.

In the process of analyzing the effectiveness of the proposed methods (achieving the aforementioned 90% accuracy of tests that will find error), the reason for this result became known — the size of the cluster that contained similar software versions averaged 80% of the total population. This is presented in Figure 2. Thus, most of the bugs were found due to a high coverage of past software versions. This means two things. First, daily software changes are rather uniform in terms of frequency and magnitude of changes in each software component. Second, most bugs are introduced in the components that change most frequently.

This fact is confirmed by the metrics runs of Silhouette and Davies-Bouldin. For the hierarchical (agglomerative) clustering and k-means methods, the average score of the first is close to 0.9, so in theory the samples are separated as perfectly as possible. The results of the second metric, which oscillate in values of 0.2 to 0.5, also report this, indicating that the ratio between cluster scatter and the cluster's separation is rather good. The case is slightly different for Gaussian mixtures, which, unlike the previous two methods, does not calculate the similarity of samples based on the distance metric, but on their correlation. In this case, we see a much greater discrepancy of these results, and even in the context of each successive version of the software — by stating changes in the code as a uniform — successive samples may have different correlation coefficients, which makes it difficult in this method to find the optimal number of clusters. Indeed, it is noticeable that the results are much lower for Silhouette, which is around 0.2, and for DB, which ranges from 1 to 3.



Rysunek 2: Cluster sizes for each selected clustering method



Rysunek 3: Average value of wrong predictions for each method (depending on the number of clusters) grouped into sprints (feature builds)

The average number of classification errors is not closely related to the number of clusters. It further confirms the fact that most errors occur in a certain group of frequently changed components. By adjusting this parameter (e.g., increasing the number of clusters), we do not risk a significant average decrease in the efficiency of test classification, but we get the possibility of a significant increase in the value of the test set optimization metric. Thus, the conclusion is that the majority of mistakes are made by the methods in the case of software changes in new or rarely changed components, and the proposed methods can serve rather to roughly approximate the areas of software where, due to similarity to past changes, they have a high probability of having bugs in them.

6. CONCLUSION AND GUIDELINES

One of the features of the historical approach is the ability to observe that over a selected period of time (here: 180 days), it is on average about half of the test set that finds errors. It can be seen that other half in this time range never fail (these are so-called 'ever green / passing' tests). An additional and undeniable advantage of the historical approach is its simplicity and ease of implementation. However, the more frequent mistakes of this approach are noticeable, which can be observed in Figure 3. The average wrong prediction ratio (in percents) are marked on the y-axis, in relation to successive software development cycles (e.g. sprints) marked on x-axis. Methods using clustering methods make similar mistakes — often in the same software versions but noticeably less often. This is due to the fact that the errors that occur at that time have not occurred before, neither in the closest history (in the last 30 or 180 days) nor in the software versions that the clustering methods found to be similar to each other. This is one of the most important drawbacks of the presented solution, which has a very limited way of finding unique and one-time bugs in software. However, the smaller number of mistakes of all clustering methods allows one to make the conclusion that the logic of using knowledge of the dynamics of changes in software components, allows in some cases to correctly predict possible errors in the code.

Therefore, our proposition is to incorporate additional information into the process. That is the knowledge related to the test contents and details on files that were changed by a recent commit (e.g. names of updated files). Such an approach will transfer the problem into a field of natural language processing, in which much more context may be extracted. As code changes were defined to be rather uniform in nature, the context of code commit (thus: description of changes and its relation to content of tests that failed) seems to be a natural extension of the dataset with additional and important content.

LITERATURA

[1] Rothermel, G. and Untch, R.H. and Chengyun Chu and Harrold, M.J. 2001. „Prioritizing test cases for regression testing". IEEE Transactions on Software Engineering, 27 (10): 929-948.

[2] João F. S. Ouriques, Emanuela G. Cartaxo, Patrícia D. L. Machado, Francisco G. O. Neto, and Ana E. V. B. Coutinho. 2016. „On the Use of Fault Abstractions for Assessing System Test Case Prioritization Techniques". In Proceedings of the 1st Brazilian Symposium on Systematic and Automated Software Testing (SAST). 7: 1–10.

[3] Tanzeem Bin Noor and Hadi Hemmati. 2017. „Studying test case failure prediction for test case prioritization". In Proceedings of the 13th International Conference on Predictive Models and Data Analytics in Software Engineering: 2–11.

[4] Francis Palma, Tamer Abdou, Ayse Bener, John Maidens, and Stella Liu. 2018. „An improvement to test case failure prediction in the context of test case prioritization". In Proceedings of the 14th International Conference on Predictive Models and Data Analytics in Software Engineering: 80–89.

[5] Ajmer Singh, Rajesh Kumar Bhatia, and Anita Singhrova. 2019. „Machine learning based test case prioritization in object oriented testing". International Journal of Recent Technology and Engineering, 8 (3): 700–707.

[6] Junjie Chen, Yiling Lou, Lingming Zhang, Jianyi Zhou, Xiaoleng Wang, Dan Hao, and Lu Zhang. 2018. „Optimizing test prioritization via test distribution analysis". In Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering: 656–667.

[7] Muhammad Hasnain, Muhammad Fermi Pasha, Chern Hong Lim, and Imran Ghan. 2019. „Recurrent neural network for web services performance forecasting, ranking and regression testing". In 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC): 96–105

[8] Siavash Mirarab and Ladan Tahvildari. 2008. „An empirical study on bayesian network-based approach for test case prioritization". In 2008 1st International Conference on Software Testing, Verification, and Validation: 278–287.

[9] Songyu Chen, Zhenyu Chen, Zhihong Zhao, Baowen Xu, and Yang Feng. 2011. „Using semi-supervised clustering to improve regression test selection techniques". In Fourth IEEE International Conference on Software Testing, Verification and Validation: 1–10.

[10] Shin Yoo, Mark Harman, Paolo Tonella, and Angelo Susi. 2009. „Clustering test cases to achieve effective and scalable prioritisation incorporating expert knowledge". In Proceedings of the eighteenth international symposium on Software testing and analysis (ISSTA '09): 201–212.

[11] Rafig Almaghairbe and Marc Roper. 2017. „Separating passing and failing test executions by clustering anomalies". Software Quality Journal, 25 (3): 803–840.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

WYKORZYSTANIE INFORMACJI NA TEMAT AKTYWNOŚCI UŻYTKOWNIKA DO ZWIĘKSZENIA DOKŁADNOŚCI W ULTRASZEROKOPASMOWYM SYSTEMIE LOKALIZACYJNYM

USING INFORMATION ON USER'S ACTIVITY TO IMPROVE ACCURACY IN AN ULTRA-WIDEBAND POSITIONING SYSTEM

Krystian Mandecki[1]; Marcin Kołakowski[2]

Instytut Radioelektroniki i Technik Multimedialnych, Politechnika Warszawska, Warszawa
[1] krystian.mandecki.stud@pw.edu.pl [2] marcin.kolakowski@pw.edu.pl

DOI: 10.15199/59.2023.4.45

Streszczenie: W referacie zaprezentowano metodę lokalizacji przeznaczoną do wykorzystania w ultraszerokopasmowym systemie lokalizacyjnym. Metoda, oprócz prowadzenia pomiarów w interfejsie radiowym, zakłada ciągłe monitorowanie aktywności użytkownika co, w połączeniu z informacją na temat rozkładu pomieszczenia, pozwala uzyskać dodatkowy kontekst. W zaproponowanej metodzie, użytkownik jest lokalizowany za pomocą filtru cząsteczkowego. Dodatkowy kontekst jest wykorzystywany na etapie ponownego próbkowania poprzez zwiększenie wag cząsteczek znajdujących się w obszarach, w których można wykonywać wykrytą aktywność.

Abstract: The paper presents a localization method intended for use in an ultra-wideband positioning system. The method, besides performing measurements in the radio interface, monitors the user's activity, which combined with the indoor area layout results in an additional positioning context. In the proposed method, the user is located using a particle filter. The additional context is used to increase the weights of particles located in the areas, where the detected activity can be performed.

Słowa kluczowe: detekcja aktywności, lokalizacja, uczenie głębokie, UWB

Keywords: activity detection, deep learning, positioning, UWB

1. WPROWADZENIE

W ostatnich latach obserwowany jest znaczący wzrost liczby zastosowań radiowych systemów lokalizacyjnych. Na rynku można znaleźć wiele rozwiązań pozwalających na świadczenie różnych usług opartych na lokalizacji np. nawigacji w centrach handlowych lub monitorowaniu procesu produkcji. Poprawna realizacja usług wymaga zapewnienia wyników lokalizacji o wystarczającej dokładności.

Dokładność wyników lokalizacji zależy od kilku czynników, z którym głównym jest technika wykorzystywana przez system. W systemach bazujących na pomiarach poziomu sygnału w interfejsach WiFi lub Bluetooth typowa dokładność jest rzędu pojedynczych metrów. Dużo lepsze wyniki zapewnia wykorzystanie techniki ultraszerokopasmowej, w której mierzone są czasy nadejścia sygnałów. Pozwala ona na uzyskanie dokładności na poziomie kilkunastu centymetrów.

Drugim istotnym czynnikiem są warunki propagacyjne, w których pracuje system. Największe dokładności uzyskiwane są w warunkach LoS (ang. Line of Sight), w których bezpośrednia ścieżka między lokalizowaną etykietą a węzłami infrastruktury nie jest przysłonięta żadnymi przeszkodami. Niestety, zapewnienie takich warunków pracy w środowiskach wewnątrzbudynkowych jest niezwykle trudne ze względu na dużą liczbę przeszkód oraz ograniczoną liczbę węzłów infrastruktury. W typowych scenariuszach system przez większość czasu pracuje w warunkach NLoS (ang. Non Line of Sight). Problem ten jest szczególnie widoczny w systemach lokalizacji osób, w których ciało użytkownika wprowadza dodatkowe tłumienie i opóźnienie powodujące błędy lokalizacji na poziomie pojedynczych metrów [1].

Jednym ze sposobów redukcji błędów lokalizacji użytkownika jest zastosowanie algorytmów uwzględniających informacje o kontekście lokalizacji. W typowych rozwiązaniach algorytm dokonuje klasyfikacji aktywności użytkownika i wykorzystuje tą wiedzę w celu poprawy dokładności. Przykładowo, wykrycie wchodzenia po schodach [2] lub otwierania drzwi [3] powoduje przyciągnięcie wyników lokalizacji odpowiednio do obszaru klatki schodowej lub przejścia między pomieszczeniami.

Detekcja aktywności użytkownika może być przeprowadzana na podstawie danych otrzymanych z noszonych czujników inercyjnych [4]. Do pozyskania informacji dotyczącej wykonywanej przez człowieka aktywności stosuje się różne podejścia. Jednym z nich jest zastosowanie modeli uczenia maszynowego takich jak drzewa decyzyjne czy maszyny wektorów nośnych uczonych na wstępnie przetworzonych danych [5, 6]. Okazuje się jednak, że odpowiednie wytrenowanie głębokiej sieci neuronowej nawet bez wstępnego przetworzenia danych może skuteczniej wykrywać aktywność człowieka [7, 8]. Szczególnie skuteczne okazuje się zastosowanie głębokich splotowych sieci neuronowych [9, 10] lub głębokich rekurencyjnych sieci neuronowych [11].

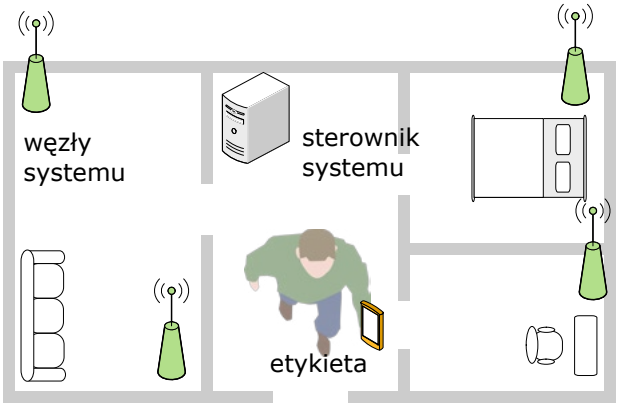
W referacie zaproponowano metodę lokalizacji, w której uwzględniany jest typ aktywności wykonywanej przez użytkownika (pozycja siedząca lub marsz). Do detekcji aktywności zastosowano splotową sieć neuronową przetwarzającą dane z trzyosiowego akcelerometru. Infor-

macja na temat aktywności jest wykorzystywana w algorytmie filtru do zwiększenia wag cząsteczek znajdujących się w obszarze, w którym można wykonywać wykrytą aktywność.

2. KONCEPCJA METODY

2.1 Architektura systemu

Schemat funkcjonalny architektury systemu, w którym może zostać zastosowana zaproponowana metoda został przedstawiony na Rys. 1.



Rysunek 1: Architektura funkcjonalna systemu lokalizacyjnego

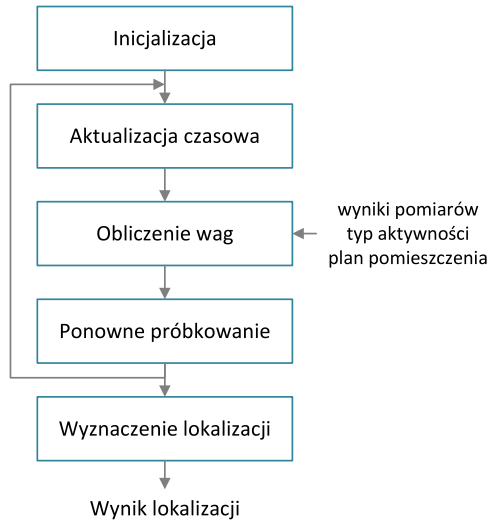
System składa się z trzech części: etykiety, węzłów oraz sterownika systemu. Etykieta jest niewielkim urządzeniem noszonym przez użytkownika. W zaproponowanym systemie pełni ona rolę źródła sygnału i okresowo nadaje pakiety UWB, które są odbierane przez węzły infrastruktury. Węzły mierzą czasy przybycia pakietów i przesyłają wyniki do sterownika systemu, w którym wyznaczana jest lokalizacja użytkownika.

W zaproponowanym rozwiązaniu etykieta jest dodatkowo wyposażona w akcelerometr, którego wskazania są wykorzystywane do określania typu aktywności użytkownika (np. marsz, pozycja siedząca). Docelowo, klasyfikacja aktywności przeprowadzana będzie za pomocą modelu uczenia maszynowego zaimplementowanego w etykiecie a jej wyniki umieszczane w polu danych pakietów UWB. W obecnej implementacji przetwarzanie jest realizowane po stronie sterownika systemu.

Dane na temat aktywności, w połączeniu z planem pomieszczenia, pozwalają na uzyskanie dodatkowych informacji na temat kontekstu lokalizacji i zwiększenie jej dokładności. Przykładowo, klasyfikacja aktywności użytkownika znajdującego się w sypialni jako leżenie wskazuje, że znajduje się on w łóżku. Informacja na temat kontekstu lokalizacji jest wykorzystywana w zaproponowanym algorytmie opartym na filtrze cząsteczkowym.

2.2 Algorytm lokalizacyjny

Zaproponowany algorytm lokalizacyjny wykorzystuje algorytm z grupy filtrów cząsteczkowych PF (ang. Particle Filters). Główną ideą algorytmów PF jest estymacja rozkładu prawdopodobieństwa zmiennej losowej (w tym przypadku lokalizacji użytkownika) na podstawie danych empirycznych (pomiarów wykonywanych przez system). Poszukiwany rozkład jest modelowany za pomocą cząsteczek, którym przypisane są wagi odpowiadające prawdopodobieństwu ich wystąpienia. Schemat blokowy zaproponowanego algorytmu przedstawiono na Rys. 2.



Rysunek 2: Schemat blokowy algorytmu lokalizacyjnego

Działanie filtru cząsteczkowego rozpoczyna się od losowej inicjalizacji N cząsteczek, z których każda opisana jest wektorem opisującym stan, w którym chwili k znajduje się użytkownik:

$$\boldsymbol{x}_{\boldsymbol{k}}^{(\boldsymbol{i})} = \begin{bmatrix} x & v_x & y & v_y \end{bmatrix} \tag{1}$$

gdzie x, y to współrzędne użytkownika, a v_x, v_y są składowymi jego wektora prędkości. Cząsteczki generowane są w sposób losowy w obszarze objętym działaniem systemu z założeniem rozkładu jednostajnego. Każdej cząsteczce przypisana jest waga w_i odpowiadająca prawdopodobieństwu jej wystąpienia.

Kolejnym etapem przetwarzania danych jest aktualizacja czasowa, w której stan użytkownika w obecnym momencie k jest estymowany na podstawie przyjętego modelu ruchu.

$$\boldsymbol{x}_{\boldsymbol{k}}^{(\boldsymbol{i})} = \boldsymbol{F}\boldsymbol{x}_{\boldsymbol{k-1}}^{(\boldsymbol{i})} + \boldsymbol{n}_{\boldsymbol{k-1}} \tag{2}$$

gdzie $\boldsymbol{F}$ jest macierzą zawierającą równania ruchu (w implementacji przyjęto model ruchu jednostajnego) a $\boldsymbol{n}_{\boldsymbol{k-1}}$ dodatkowym zaburzeniem o rozkładzie normalnym.

Wagi cząsteczek są uaktualniane na podstawie wyników pomiarów i detekcji aktywności.

$$w_i = \eta(w_{m,i} + w_{a,i})w_{g,i} \tag{3}$$

$$w_{m,i} = 1/\left(c\sqrt{\frac{\sum_n (z_{k,n} - h_{k,n}(\boldsymbol{x}_{\boldsymbol{k}}^{(\boldsymbol{i})}))^2}{\mathrm{len}(\boldsymbol{z_k})}}\right) \tag{4}$$

$$w_{a,i} = 1/d(\boldsymbol{x}_{\boldsymbol{k}}^{(\boldsymbol{i})}) \tag{5}$$

$$w_{g,i} = \begin{cases} 1, & \text{jeśli } \boldsymbol{x}_{\boldsymbol{k}}^{(\boldsymbol{i})} \text{ w granicach pomieszczenia} \\ 0, & \text{w przeciwnym przypadku} \end{cases} \tag{6}$$

Na wagę przypisywaną cząsteczce składają się trzy czynniki związane z wynikami pomiarów ($w_{m,i}$), kontekstem lokalizacji $w_{a,i}$ oraz rozkładem pomieszczenia $w_{g,i}$. Wagi są skalowane za pomocą współczynnika η tak by wagi wszystkich cząsteczek sumowały się do jedności.

Czynnik związany z wynikami pomiarów (zapisanymi w wektorze pomiarowym $\boldsymbol{z_k}$) jest obliczany jako pierwiastek średniej kwadratów różnicy między pomiarami zwracanymi przez system $z_{k,n}$ a odpowiadającymi im wartościami estymowanymi dla lokalizacji danej cząsteczki $h_{k,n}(\boldsymbol{x_k^{(i)}})$. Otrzymana wartość jest skalowana poprzez przemnożenie przez prędkość światła c by dopasować jednostki między poszczególnych czynników.

Kontekst lokalizacji jest uwzględniany poprzez dodanie czynnika $w_{a,i}$ będącego odwrotnością dystansu do najbliższego obszaru, w którym możliwe jest wykonywanie wykrytej aktywności. W przypadku detekcji aktywności, do których nie są przypisane żadne szczególne obszary czynnik ten wynosi 0.

W niektórych przypadkach, część cząsteczek może znaleźć się poza obszarem objętym działaniem systemu (np. za ścianami zewnętrznymi budynków). Aby uniknąć błędów, takim cząsteczkom przypisywana jest zerowa waga poprzez przemnożenie przez $w_{g,i}$.

Wartości cząsteczek wraz z wagami reprezentują rozkład poszukiwanej zmiennej losowej. Ostatnim krokiem przetwarzania danych jest ponowne próbkowanie polegające na celu zastąpienie cząsteczek nowymi, lepiej odpowiadającymi otrzymanemu rozkładowi. Otrzymane cząsteczki są wykorzystywane w kolejnej iteracji algorytmu.

Lokalizacja użytkownika obliczana jest jako średnia ważona cząsteczek:

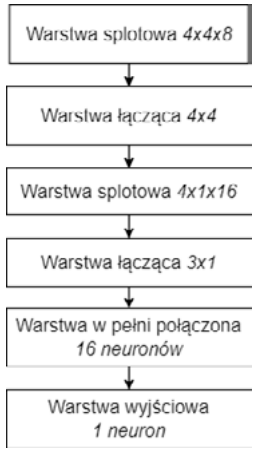
$$\boldsymbol{x_k} = \frac{\sum_i w_i \boldsymbol{x_k^{(i)}}}{\sum_i w_i} \tag{7}$$

2.3 Moduł detekcji aktywności

Noszony przez człowieka akcelerometr odczytuje wartości przyspieszeń na każdej z osi 50 razy na sekundę. Ponadto, dla każdych 3 odczytów przyspieszeń a_x, a_y, a_z w jednej chwili czasowej, liczony jest moduł przyspieszenia r z następującego wzoru:

$$r = \sqrt{a_x^2 + a_y^2 + a_z^2} \tag{8}$$

Detekcja aktywności jest przeprowadzana za pomocą sieci splotowej na podstawie przyspieszeń a_x, a_y, a_z, r otrzymanych w ciągu ostatnich 2 sekund. Architekturę zastosowanej sieci splotowej przedstawiono na Rys. 3.



Rysunek 3: Architektura sieci neuronowej

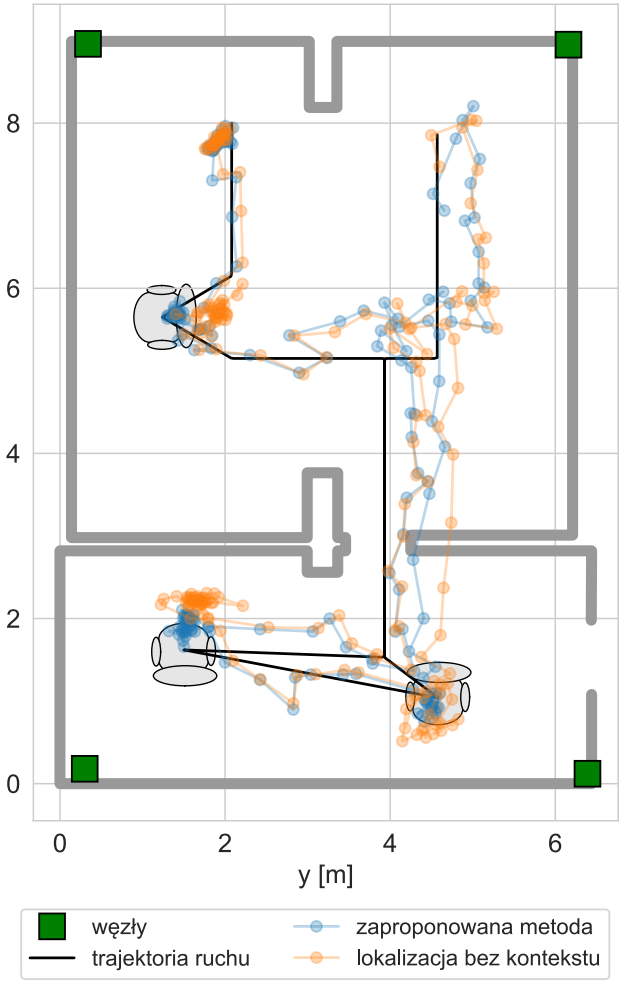
W zaproponowanym modelu sieci neuronowej, za ekstrakcję cech odpowiadają dwie warstwy splotowe, po każdej z nich zastosowano maksymalizującą warstwę łączącą. Przed warstwą wyjściową zastosowano warstwę w pełni połączoną.

W każdej warstwie łączącej i w warstwie w pełni połączonej zastosowano porzucanie. Przyjęto współczynnik porzucania równy 0,1. Użyta funkcja aktywacji w warstwach splotowych oraz warstwie w pełni połączonej to ReLU. Warstwę wyjściową stanowi jeden neuron z sigmoidalną funkcją aktywacji. Model został zaimplementowany przy użyciu języka Python oraz biblioteki TensorFlow.

3. BADANIA EKSPERYMENTALNE

Zaproponowana metoda została zbadana eksperymentalnie w jednym z laboratoriów Instytutu Radioelektroniki i Technik Multimedialnych Politechniki Warszawskiej. W trakcie badań wykorzystano ultraszerokopasmowy system wykorzystujący technikę Time Difference of Arrival (TDoA) [12]. Plan laboratorium, wraz z lokalizacją węzłów został przedstawiony na Rys. 4.

Zastosowany system składał się z czterech węzłów rozmieszczonych na ścianach pomieszczenia oraz etykiety. Do pomiarów przyspieszenia, został wykorzystana aplikacja mobilna. W trakcie badań, zarówno telefon jak i smartfon były noszone w kieszeni. Sterownik systemu oraz smartfon były podłączone do internetu co pozwoliło na synchronizację pomiarów zwracanych przez smartfon i infrastrukturę.



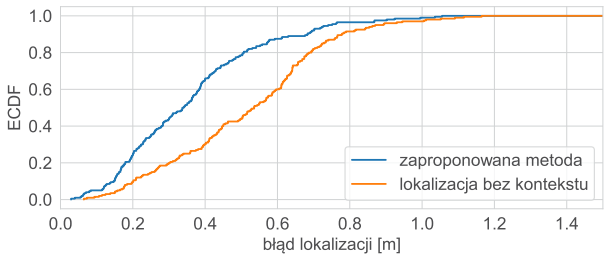
Rysunek 4: Rozkład pomieszczenia i przykładowe wyniki lokalizacji

Eksperyment polegał na lokalizacji osoby przemieszczającej się po zdefiniowanej trajektorii i siadającej na trzech krzesłach zlokalizowanych w pomieszczeniu. Podczas eksperymentu, rejestrowane były czasy, w których osoba przechodziła przez kolejne wierzchołki trajektorii oraz siadała. Pozwoliło to na uzyskanie referencyjnych lokalizacji użytkownika (założono, że osoba poruszała się ze stałą prędkością wzdłuż poszczególnych odcinków trajektorii), które zostały wykorzystane do obliczenia błędu lokalizacji.

Osoba lokalizowana była za pomocą zaproponowanej metody oraz konwencjonalnego filtru cząsteczkowego, w którym kontekst lokalizacji nie był uwzględniany (waga $w_{a,i}$ zawsze była równa 0). Otrzymane wyniki zostały przedstawione na Rys. 4.

Wyniki uzyskane za pomocą zaproponowanej metody cechują się większą jakością niż w przypadku wykorzystania konwencjonalnego filtru cząsteczkowego. Przypisanie większych wag cząsteczkom znajdującym się w pobliżu krzesła w momentach gdy aktywność osoby klasyfikowana jest jako siedzenie pozwala na zwiększenie dokładności lokalizacji. Dokładniejsza estymacja rozkładu wektora stanu pozytywnie wpływa na jakość kolejnych wyników. Otrzymana trajektoria ruchu jest bliżej krzywej referencyjnej.

Dokładniejsze porównanie metod można przeprowadzić na podstawie błędu lokalizacji. Empiryczne dystrybuanty błędów zostały przedstawione na Rys. 5. W przypadku obydwu testowanych metod, maksymalny błąd lokalizacji nie przekroczył 1,1 m. Dla zaproponowanej metody, średni błąd wynosi 0,36 m i jest znacznie mniejszy niż w przypadku zastosowania konwencjonalnego filtru cząsteczkowego (0,52 m).



Rysunek 5: Empiryczne dystrybuanty błędu lokalizacji

4. PODSUMOWANIE

W referacie przedstawiono metodę lokalizacji przeznaczoną do zastosowania w ultraszerokopasmowych systemach lokalizacyjnych. Metoda, oprócz pomiarów wykonywanych w interfejsie radiowym wykorzystuje informacje dotyczące aktywności wykonywanej przez użytkownika. Detekcja aktywności jest przeprowadzana na podstawie pomiarów przyspieszenia za pomocą klasyfikatora opartego na splotowej sieci neuronowej.

Wyniki badań eksperymentalnych pokazały, że uwzględnienie dodatkowej informacji na temat kontekstu lokalizacji pozwala na zwiększenie dokładności wyznaczania położenia użytkownika. Średni błąd lokalizacji w porównaniu do zastosowania konwencjonalnego filtru cząsteczkowego był mniejszy o kilkanaście centymetrów.

Skuteczność zaproponowanej metody można dodatkowo zwiększyć poprzez uwzględnienie innych typów aktywności np. wchodzenie po schodach lub otwieranie drzwi. Nie przeprowadzono jeszcze badań w tym zakresie.

W referacie przedstawiono wyniki prac realizowanych w ramach projektu finansowanego przez Narodowe Centrum Badań i Rozwoju w ramach programu AAL Call 2021 nr umowy AAL/AALCall2021/52/CAREUP/2022.

LITERATURA

[1] M. Kołakowski, 2020, “Redukcja wpływu ciała użytkownika na dokładność lokalizacji z wykorzystaniem hybrydowego algorytmu lokalizacyjnego,” Przegląd Telekomunikacyjny - Wiadomości Telekomunikacyjne (7-8), 312–315, DOI: 10.15199/59.2020.7-8.39

[2] G. Pipelidis i inni, 2019 “A Novel Lightweight Particle Filter for Indoor Localization,” 2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN), (Pisa, Włochy), 1–8, DOI: 10.1109/IPIN.2019.8911744

[3] R. Jackermeier i B. Ludwig, 2019, “Door Transition Detection for Long-Term Stability in Pedestrian Indoor Positioning,” 2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN), (Pisa, Włochy), 1–8, DOI: 10.1109/IPIN.2019.8911810.

[4] Zawar Hussain, Michael Sheng, Wei Emma Zhang, 2019. “Different Approaches for Human Activity Recognition: A Survey,” DOI: 10.48550/arXiv.1906.05074.

[5] Akram Bayat, 2014, “A Study on Human Activity Recognition Using Accelerometer Data from Smartphones,” DOI: 10.1016/j.procs.2014.07.009.

[6] Oscar D. Lara, Miguel A. Labrador, 2013, “A survey on human activity recognition using wearable sensors,” DOI: 10.1109/SURV.2012.110112.00192.

[7] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, Lisha Hu, 2019. “Deep learning for sensor-based activity recognition: A survey,” DOI: 10.1016/j.patrec.2018.02.010.

[8] Anna Ferrari, Daniela Micucci, Marco Mobilio, Paolo Napoletano, 2019, “Human Activities Recognition Using Accelerometer and Gyroscope,” DOI: 10.1007/978-3-030-34255-5_28.

[9] Wenchao Jiang, Zhaozheng Yin, 2015 “Human Activity Recognition Using Wearable Sensors by Deep Convolutional Neural Networks,” DOI: 10.1145/2733373.2806333.

[10] Charissa Ann Ronao, Sung-Bae Cho, 2016 “Human activity recognition with smartphone sensors using deep learning neural networks,” DOI: 10.1016/j.eswa.2016.04.032.

[11] Abdulmajid Murad, Jae-Young Pyun, 2017 “Deep Recurrent Neural Networks for Human Activity Recognition,” DOI: 10.3390/s17112556.

[12] J. Kołakowski i inni, 2020, “UWB/BLE Tracking System for Elderly People Monitoring,” Sensors, 20, 1574, DOI: 10.3390/s20061574

ZARZĄDZANIE POLITYKĄ W MOBILNEJ SIECI RADIA KOGNITYWNEGO Z DYNAMICZNYM DOSTĘPEM DO WIDMA

POLICY MANAGEMENT IN MOBILE COGNITIVE RADIO NETWORK WITH DYNAMIC SPECTRUM ACCESS

Piotr Gajewski[1]; Jerzy Łopatka[2]; Radosław Chęciński[3], Anna Kaszuba-Chęcińska[4]

[1] Wydział Elektroniki Wojskowa Akademia Techniczna Warszawa, piotr.gajewski@wat.edul.pl
[2] Wydział Elektroniki Wojskowa Akademia Techniczna Warszawa, jerzy.lopatka@wat.edul.pl
[3] Wydział Elektroniki Wojskowa Akademia Techniczna Warszawa, radoslaw.checinski@wat.edul.pl
[4] Wydział Elektroniki Wojskowa Akademia Techniczna Warszawa, anna.kaszuba@wat.edul.pl
…

DOI: 10.15199/59.2023.4.46

Streszczenie: **Polityki radiowe stają się znaczącym składnikiem systemów bezprzewodowych nowych generacji wymagającym efektywnych narzędzi zarządzania z zastosowaniem sztucznej inteligencji i uczenia maszynowego. W artykule przedstawiono proces zarządzania polityką w mobilnym systemie ad-hoc (MANET) na podstawie oceny strategii jego funkcjonowania dla różnorodnych przypadków użycia. Wykazano przydatność metody hierarchicznej dekompozycji decyzji dla wielowymiarowego procesu oceny strategii dynamicznego zarządzania widmem w systemie MANET z węzłami kognitywnymi.**

Abstract: **Radio policies are becoming a significant component of next-generation wireless systems, requiring effective management tools, using artificial intelligence and machine learning. The article presents the process of policy management in mobile ad hoc network (MANET) based on evaluation of the system strategy for various use cases. The usefulness of the hierarchical decision decomposition method for the multidimensional process of the dynamic spectrum management strategy evaluation in the MANET system with cognitive nodes has been shown.**

Słowa kluczowe: **Radio Kognitywne, Dynamiczne Zarządzanie Widmem, Polityki Radiowe, Zarządzanie Polityką w MANET.**

Keywords: **Cognitive Radio, Dynamic Spectrum Management, Radio Policies, Policy Management in MANET.**

1. WSTĘP

Jednym z paradygmatów rozwoju systemów radiokomunikacyjnych jest coraz szersze wykorzystanie sztucznej inteligencji AI (Artificial Intelligence) z uczeniem maszynowym ML (Machine Learning), zarówno w obszarach przetwarzania sygnałów, jak i zarządzania [1], [2]. W heterogenicznych sieciach następnych generacji NG (6G/7G i dalszych) przewidywana jest kongruencja sieci dostępowych i doraźnych (ad hoc), a ciągły rozwój obszarów zastosowania NG znakomicie rozszerza liczbę i zakres scenariuszy dywersyfikujących klasy usług sieci, zarówno komercyjnych, jak i specjalnych, w tym wojskowych sieci mobilnych ad hoc z węzłami kognitywnymi CR (Cognitive Radio).

W ramach prac związanych z wykorzystaniem CR w sieciach wojskowych opracowana została koncepcja radia XG (neXt Generation) [3], w której założono wykorzystanie tej technologii w złożonym środowisku o zróżnicowanych wymaganiach, zapisanych jako polityki wykorzystywane w zarządzaniu sieciami. Polityki te definiują mechanizmy konfigurowania i zachowania oprogramowania w sposób, który zwykle nie jest przewidywalny podczas definiowania polityki [4].

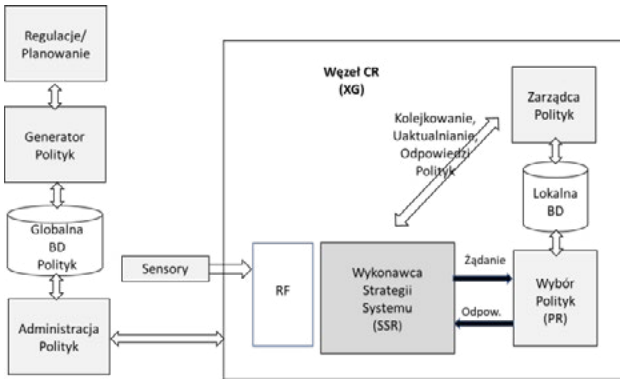
Prowadzone są intensywne badania związane z wdrażaniem nowych technologii, gdzie inteligentne zarządzanie aplikacjami wymaga stosowania odpowiednich strategii wyboru sposobu realizacji funkcji systemu i doboru odpowiednich polityk z tym związanych. W niniejszym opracowaniu przedstawiono propozycję zastosowania metody AHP (Analytic Hierarchy Process) [5], do rozwiązania problemu oceny i wyboru strategii w taktycznej sieci MANET (Mobile Ad hoc Network) z węzłami kognitywnymi, w której stosowane jest dynamiczne zarządzanie dostępem do widma DSM (Dynamic Spectrum Management). W [6] autorzy opisali model systemu opracowany w ramach projektu MAENA (Multiband Efficient Network for ad hoc communications), gdzie wykorzystano mieszany sposób zarządzania widmem z zastosowaniem procedur zarządzania scentralizowanego C-DSM (Central DSM) i rozproszonego L-DSM (Local DSM). Wyniki oceny właściwości takiego systemu dla scenariuszy z włączonymi i wyłączonymi mechanizmami DSM dla wybranego scenariusza operacyjnego przedstawiono w [7].

W rozdziale 2 opisano koncepcję i rozwiązania strukturalne radia bazującego na politykach PBR (Policy Based Radio), przedstawiono ogólną charakterystykę polityk i sposobów zarządzania wykorzystującymi je strategiami w sieci MANET CR. Ocena i wybór strategii jest kluczowa dla efektywnego funkcjonowania sieci w warunkach szybko zmieniającego się otoczenia. Przykład wykorzystania do tej oceny zmodyfikowanej metody AHP w taktycznym systemie radiowym dla wybranych scenariuszy, typowych dla działań operacyjnych, opisany został w rozdziale 3 wraz z wynikami oceny. Rozdział 4 zawiera wnioski z propozycją rozwoju w symulatorze MAENA.

2. ZARZĄDZANIE W SIECIACH RADIA XG

2.1. Rozwiązania strukturalne w MANET CR

Koncepcja systemu CR opartego na politykach przedstawiona jest na Rys.1. Węzeł CR realizuje politykę wybraną z lokalnej bazy według reguł określonych przez PR (Policy Reasoner), zgodnie z żądaniem przekazanym z bloku wykonawcy strategii systemu SSR (System Strategy Reasoner). Polityki do lokalnej BD są pobierane, porządkowane i uaktualniane przez zarządcę z systemu centralnej dystrybucji.



Rys.1. Szkielet systemu PBR

Polityki mogą dotyczyć różnych funkcjonalności systemu. W niniejszym opracowaniu są one związane z dynamicznym zarządzaniem widmem DSM w sieci MANET CR [6], a zatem obejmują zasady: eksploracji widma, eksploatacji widma, dystrybucji informacji widmowych, a w systemie kognitywnym również uczenie zachowania.

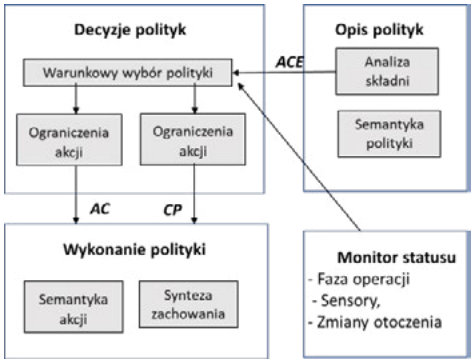
W tym rozwiązaniu polityki wybierane z globalnej bazy danych są dystrybuowane centralnie przez Administratora Polityk i zarządzane indywidualnie z lokalnej BD w węzłach CR, w których tzw. silnik polityk współdziała z silnikiem kognitywnym. Ogólnie, centralne zarządzanie obejmuje: planowanie polityk z wyborem regulacji, generowanie polityk i administrowanie politykami. Zarządzanie lokalne polega na wyborze strategii DSM najskuteczniejszej z punktu widzenia osiągnięcia celu, ustalonego dla konkretnej misji. W artykule [8] przedstawiono koncepcję standaryzacji architektury hierarchicznego Dynamicznego Współdzielenia Widma (H-DSS), która wykorzystuje nową, opartą na polityce misji metodę DSA z implementacją narzędzi do zarządzania widmem obejmujących: definiowanie, implementację i ocenę polityk.

2.2. Polityki radiowe w sieciach MANET CR

Polityki DSM są zwykle definiowane jako mechanizmy do definiowania i kodowania w przyjętym języku polityki specyficznych reguł DSA (np. kanały radiowe, poziom mocy, sposób nadawania,…), dla określonych warunków funkcjonowania systemu zgodnych z wcześniej ustalonymi zasadami i sposobami zmiany zachowania systemu przy realizacji operacji [9], [10].

Przy budowaniu systemu PBR uwzględnia się dwie fazy: kreowania polityki na podstawie strategii działania systemu oraz opracowania procedury zarządzania politykami. W [4] zdefiniowane zostały trzy kategorie polityk:

- Akcje zdarzenia warunkowego ECA (Event-Condition-Action) – specyfikujące zachowanie (akcje) jako odpowiedź na określone zdarzenia.
- Sterowanie akcjami AC (Action Control) – reguły specyfikujące czy akcja jest dozwolona, czy zabroniona.
- Polityki konfiguracyjne CP (Configuration Policies).



Rys. 2. System radiowy z PBR

Na Rys. 2 przedstawiono ogólny schemat realizacji funkcjonalności sieci z węzłami kognitywnymi opartymi na politykach PBR. Zawiera on 4 bloki: definiowania polityki (opisów), wyboru polityk (decyzji), wykonania polityk i monitora statusu. Najbardziej złożonym problemem jest wybór polityk do realizacji dla przewidywanych przypadków funkcjonowania sieci w złożonym środowisku sieci nowych generacji (wojskowych i cywilnych), gdzie baza polityk może zawierać setki, a nawet tysiące polityk. Dlatego ciągle poszukiwane są rozwiązania problemu wielokryterialnego i wielopoziomowego, w tym możliwych do wykorzystania z zastosowaniem sztucznej inteligencji z uczeniem maszynowym.

W zaproponowanym poniżej przykładzie przyjęto hybrydowy sposób zarządzania jak stosowany w projekcie MAENA dla DSM.

3. STRATEGIE I POLITYKI W SIECIACH SPECJALNYCH

3.1. Taksonomia strategii DSM

Zakładając, jak wcześniej, cztery obszary funkcjonowania DSM możemy wyodrębnić 9 przypadków (opcji) realizacji polityk, przedstawionych w Tab. 1.

Tab. 1. Opcje funkcjonalne DSM

Obszar	*Funkcja*	*Opcje*
Eksploracja widma	Monitorowanie	Scentralizowane
		Rozproszone
	Wymiana danych	Kooperacyjna
		Niekooperacyjna
Eksploatacja widma	Podejmowanie decyzji	Scentralizowane
		Rozproszone
	Sposób realizacji decyzji	Bezkonfliktowy
		Rywalizacyjny
	Użytkowanie	Kooperacyjne
		Niekooperacyjne
Dystrybucja danych	Sposób	Separacyjny
		Zintegrowany
	Rodzaj danych	Lokalne
		Globalne
Uczenie	Wykonanie	Scentralizowane
		Rozproszone
	Wykorzystanie	Kooperacyjne
		Niekooperacyjne

Z połączenia wymienionych obszarów, funkcji i opcji uzyskuje się zbiór 32 strategii DSM, z tego 20 ma znaczenie praktyczne, ponieważ 12 zawiera opcje wykluczające się logicznie.

3.2. Metryki strategii DSM

W pracy zdefiniowano zbiór metryk dla czterech podstawowych domen: świadomość środowiskowa, kognitywność, ogólna wydajność i złożoność.

Do oceny funkcji eksploracji (sensingu) czyli budowania świadomości środowiskowej przyjęto 4 metryki: geoprzestrzenne wsparcie operacji, adaptacja sensingu (wsparcie mobilności), wydajność w gęstym środowisku z zakłóceniami, efektywność gromadzenia i dystrybucji danych sensingu.

Funkcję kognitywności węzłów i sieci przyjęto oceniać za pomocą: zdolności realizacji funkcji, efektywności gromadzenia i dystrybucji danych zarządzania kognitywnego CM (Cognitive Management).

Do oceny ogólnej wydajności wykorzystano: żywotność sieci, efektywność wykorzystania widma, zdolność wsparcia szybkich zmian środowiska, zdolność adaptacji do wymagań operacyjnych (misji).

Złożoność zaproponowano oceniać przy pomocy kosztów implementacji.

3.3. Ocena i wybór strategii DSM

Ogólny szkielet definiowania polityki dostępu do widma [12] zawiera silnik budowy polityk obejmujący fazę budowy i fazę oceny. W projekcie w fazie definiowania wykorzystano metodę ekspertów, a w fazie oceny zastosowano metodę AHP [5], która ma charakter w dużym stopniu subiektywny, ale umożliwia rozwiązanie oceny problemów złożonych, wielokryterialnych. W AHP multidecyzyjny problem jest dekomponowany w zbiór wykluczających się przypadków (alternatyw) tworząc hierarchiczny model analityczny.

Do oceny strategii DSM przyjęto następującą strukturę procesu:
- określenie ocen metryk,
- wyznaczenie sumarycznej wartości dla każdej strategii,
- określenie współczynników wagowych metryk oceny strategii dla każdego scenariusza,
- ocenę każdego przypadku.

W pracy przyjęto 3 scenariusze najbardziej typowe dla działań operacyjnych:
- SC1 - Ochrona bazy (kampusu), gdzie występuje zmienne zagęszczenie wielu sieci statycznych, lub o małej mobilności;
- SC2 - Konwój (patrol), gdzie występuje kilka sieci z węzłami poruszającymi się wzdłuż wyznaczonych trajektorii, o zmiennych odległościach między drogami;
- SC3 - Walka, gdzie węzły sieci są rozproszone w środowisku zabudowanym, a praca sieci jest celowo zakłócana.

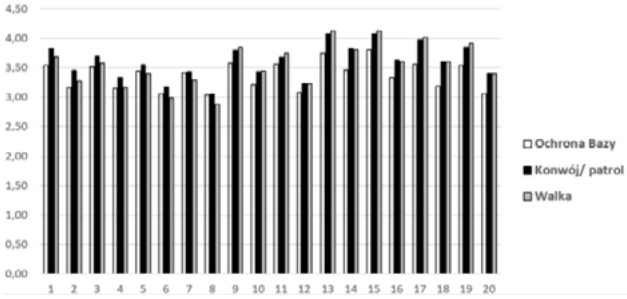
W Tab. 2 zamieszczono oceny eksperckie oraz wagi AHP poszczególnych metryk w ocenianych domenach dla każdego scenariusza. Komentując oceny:
- Geoprzestrzenne wsparcie operacji ma najwyższą ocenę dla SC1, ze względu na zagęszczenie wielu węzłów i wielu sieci;
- Adaptacja sensingu jest najistotniejsza w SC3 ze względu na dużą dynamikę zmian otoczenia i struktury sieci;
- Wydajność w gęstym i zakłóconym środowisku ma wysokie znaczenie w SC1, ze względu na gęstość węzłów i w SC3, ze względu na zakłócenia;
- Efektywność gromadzenia i dystrybucji danych sensingu dla SC1 i SC2 jest istotna z uwagi na koncentrację węzłów;
- Zdolności funkcji kognitywnych oraz efektywność gromadzenia i dystrybucji danych CM mają największe znaczenia w SC3 z uwagi na dynamikę zmian położenia węzłów i unikanie zakłóceń;
- Żywotność sieci ma priorytet w każdym scenariuszu;
- Efektywność wykorzystania widma ma najwyższą ocenę dla SC1 i SC3 z uwagi na duże zapotrzebowanie na kanały łączności;
- Zdolność wsparcia szybkich zmian środowisk oraz zdolność adaptacji do wymagań operacji są podstawowe w SC3;
- Koszt implementacji jest najwyższy w SC1.

Tab. 2. Oceny i wagi metryk dla scenariuszy

Domena	**Ocena**			**Waga AHP**		
Metryka	SC1	SC2	SC3	SC1	SC2	SC3
Świadomość środowiskowa (sensingu)						
Geoprzestrzenne wsparcie operacji	5	3	2	0.108	0.075	0.042
Adaptacja sensingu (Wsparcie mobilności)	3	3	5	0.065	0.075	0.104
Wydajność w gęstym i zakłóconym środowisku	5	3	4	0.108	0.075	0.083
Efektywność gromadzenia i dystrybucji danych sensingu	5	3	5	0.108	0.075	0.104
Kognitywność						
Zdolności funkcji kognitywnych	3	4	5	0.063	0.100	0.104
Efektywność gromadzenia i dystrybucji danych CM	3	4	5	0.065	0.100	0.104
Ogólna wydajność						
Żywotność sieci	5	5	5	0.108	0.125	0.104
Efektywność wykorzystania widma	5	3	5	0.108	0.075	0.104
Zdolność wsparcia szybkich zmian środowiska	3	4	5	0.065	0.100	0.104
Złożoność implementacyjna						
Koszt implementacji	5	3	2	0.108	0.075	0.042
Wynik ogólny	**46**	**40**	**48**			

Aby ocenić 20 strategii dla 3 scenariuszy stosując 10 metryk należy wyznaczyć ponad 600 wartości współczynników wagowych.

Wyniki globalnej oceny strategii DSM dla sieci MANET CR przedstawiono na Rys. 3, a w tabeli 3 zamieszczono wyniki uzyskane dla strategii z najwyższymi ocenami.



Rys. 3. Zestawienie wyników oceny strategii DSM w sieci MANET CR

Tab. 3. Wyniki oceny AHP do wyboru strategii DSM

Scenariusz	Strategia						
	st1	st9	st13	st14	st15	st17	st19
SC1	3,54	3,58	**3,75**	3,46	**3,81**	3,56	3,54
SC2	3,83	3,80	**4,08**	**3,83**	**4,08**	**3,98**	**3,85**
SC3	3,69	3,85	**4,13**	3,81	**4,13**	**4,02**	**3,92**

Widać, że najlepsze rezultaty są osiągalne przy stosowaniu dla scenariuszy mobilnych strategii 13 i 15 (rozproszony sensingu kooperacyjny, rozproszone zarządzanie widmem z kooperacją węzłów zarówno w sytuacjach stosowania procedur unikania konfliktów, jak i rywalizacji). W scenariuszu 3 (zakłócenia aktywne) można wykorzystać strategie 17 i 19 (brak kooperacji sensingu), natomiast dla scenariusza 2 pięć strategii daje podobne rezultaty (1, 9, 13, 15, 17, 19).

4. WNIOSKI

Omówiona w artykule metodyka zarządzania politykami w oparciu o oceny dokonane przez dekompozycję strategii jest obiecująca zarówno w statycznym planowaniu użycia sieci kognitywnej MANET, jak i w trakcie realizacji misji. Wskazują na to wyniki uzyskane dla scenariuszy mobilnych (Konwój, Walka). Wybór strategii dla scenariusza statycznego (Ochrona bazy) jest jednak utrudniony z uwagi na niewielkie różnice wartości ocen uzyskanych dla tego scenariusza. W tym przypadku lepsze rezultaty może dać scentralizowane monitorowanie widma i zcentralizowane zarządzanie nim.

Przydatność metody zostanie zweryfikowana podczas badań z zastosowaniem symulatora MAENA umożliwiającego modelowanie zarówno strategii opisanych w artykule, jak i wykorzystujących metody hybrydowe łączące zarządzanie scentralizowane i rozproszone.

Przedstawiona metoda może być wykorzystana przy uczeniu algorytmów DL w zaawansowanych metodach DSM wykorzystujących procedury AI.

5. PODZIĘKOWANIA

Niniejsza praca badawcza została wykonana w ramach projektu MAENA EDA nr B-1476-IAP4-GP i oraz w ramach Projektu Badawczego Wojskowej Akademii Technicznej nr UGB/22-860/23.

LITERATURA

[1] Akyildiz I. F., Kak A. , Nie S.. "6G and Beyond: The Future of Wireless Communications Systems". 2020. *IEEE Access*. vol. 8, pp. 133995-134030. doi: 10.1109/ACCESS.2020.3010896.

[2] Pham Quoc-Viet i in.. "Intelligent Radio Signal Processing: A Survey". in 2021. *IEEE Access*. vol. 9, pp. 83818-83850. doi: 10.1109/ACCESS.2021.3087136.

[3] Wilkins David i in.. "Policy-based Cognitive Radio". 2007. *IEEE Wireless Communications,* August 2007, pp. 41-45.

[4] Chandria Ritu. "Beyond the hype: policies for military network operations". 2006. *International Conference on Systems and Networks Communication* (ICSNC'06), DOI:10.1109/ICSNC.2006.25.

[5] Omkarprasad S. Vaidya, Sushil Kumar. "Analytic hierarchy process: An overview of applications". 2006. *European Journal of Operational Research*. Volume 169. 1, pp. 1–29.

[6] Gajewski Piotr, Łopatka Jerzy, Łubkowski Piotr. "Modelowanie wielozakresowej sieci taktycznej z Dynamicznym Zarządzaniem Widmem". 2022. *Przegląd Telekomunikacyjny, Wiadomości Telekomunikacyjne nr 4*, *Multikonferencja KSTiT i KKRRiT 2022*. doi: 10.15199/59.2022.4.96.

[7] Gajewski Piotr, Łopatka Jerzy, Łubkowski Piotr. "Performance Analysis of Public Safety Cognitive Radio MANET for Diversified Traffic". 2022. *Sensors* 22(5). doi:10.3390/s22051927.

[8] Woods Timoty i in.. "All-domain spectrum command and control via hierarchical dynamic Spectrum Sharing with implemented Dynamic Spectrum Access toolchain". 2021. *2021 IEEE Military Communications Conference (MILCOM)*, San Diego, CA, USA, pp. 354-359, doi: 10.1109/MILCOM 52596.2021.9653092

[9] Jesuale Nancy, Eydt Bernard. "A Policy Proposal to Enable Cognitive Radio for Public Safety and Industry in the Land Mobile Radio Bands". 2007. *2nd IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks*. Dublin, Ireland, 2007. pp. 66-77, doi: 10.1109/DYSPAN. 2007.16.

[10] Boksiner Jeffrey, i in.. "Policy-based spectrum management architecture". 2013. *IEEE Military Communications Conference*, doi: 10.1109/ MILCOM.2012.6415720.

[11] Lewis David i in.. "Managing policies for Dynamic Spectrum Access". 2006. *Autonomic Networking. AN 2006. Lecture Notes in Computer Science*, vol 4195. Springer, Berlin, Heidelberg.

[12] Santos Henrique i in.. "The dynamic spectrum access policy framework in action". https://ceur-ws.org/Vol-2721/paper555.pdf

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

ZASTOSOWANIE SIECI NEURONOWYCH DO DETEKCJI IMPULSÓW O ZNANYM KSZTAŁCIE W OBECNOŚCI SILNEGO SZUMU I TRENDU

APPLICATION OF NEURAL NETWORKS FOR DETECTION OF PULSES OF KNOWN SHAPE IN THE PRESENCE OF STRONG NOISE AND TREND

Szymon Betka[1]; Marek Blok[2]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, szymonbetk@gmail.com
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, marek.blok@pg.edu.pl

DOI: 10.15199/59.2023.4.47

Streszczenie: **Detekcja impulsów w odebranym sygnale radiowym, zwłaszcza w obecności silnego szumu oraz trendu, jest trudnym zadaniem. Artykuł przedstawia propozycje rozwiązań wykorzystujących sieci neuronowe do detekcji impulsów o znanym kształcie w obecności silnego szumu i trendu. Na potrzeby realizacji tego zadania zaproponowano dwie architektury. W pracy przedstawiono wyniki badań wpływu kształtu impulsu, mocy zakłóceń szumowych oraz trendu obecnego w sygnałach wejściowych sieci, na skuteczność detekcji zaproponowanych rozwiązań.**

Abstract: **Detecting pulses in a received radio signal, especially in the presence of strong noise and trend, is a difficult task. The article presents proposed solutions based on neural networks for the detection of pulses of known shape in the presence of strong noise and trend. Two architectures are proposed for the purpose. The paper presents the results of the study of the influence of the pulse shape, the noise power, and the trend present in the input signals of the network on the detection performance of the proposed solutions.**

Słowa kluczowe: **detekcja impulsu, detekcja pozycji impulsu, silny szum i trend, sieć neuronowa.**

Keywords: **pulse detection, pulse position detection, high noise and trend, neural network.**

1. WSTĘP

Detekcja sygnałów impulsowych w środowisku szumowym jest bardzo ważnym zagadnieniem w dziedzinach komunikacji bezprzewodowej, monitoringu morskiego, a także bioinformatyki [8], w których zakłócenia powodują znaczący spadek stosunku mocy sygnału do mocy szumu, co istotnie utrudnia lub uniemożliwia ich detekcję. W związku z istotą problemu, na przestrzeni lat powstało bardzo wiele klasycznych rozwiązań tego problemu: metody oparte na: filtracji [2], detekcji statystycznej [4], maszynie wektorów nośnych [6], a także metody oparte na regresji liniowej.

Wśród rozwiązań już istniejących nie zabrakło również takich opierających się na sieciach neuronowych i uczeniu maszynowym. W rozwiązaniach [3,5] zastosowano CNN w połączeniu z siecią LSTM oraz sieć *deep belief network* (DBF). W [7] użyto splotową sieć neuronową w celu klasyfikacji sygnałów z kodowaniem Morsa. Wymienione rozwiązania osiągały znakomite wyniki, lecz ich głównym zadaniem było wykrywanie obecności sygnałów w szumie bez uwzględniania problemu estymacji parametrów sygnałów wejściowych.

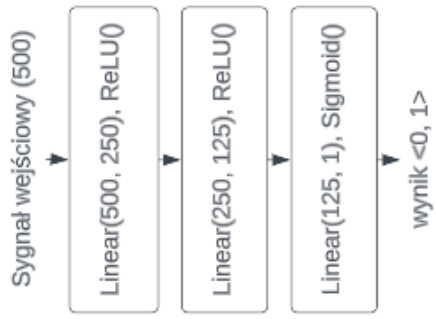
W pracy zaproponowano dwa rozwiązania realizujące detekcję impulsów o znanym kształcie w obecności silnego szumu i trendu. Pierwsze rozwiązanie realizuje detekcję bazującą na wskazywaniu obecności impulsu w sygnale wejściowym, a drugie skupia się na detekcji łączącej się z estymacją podstawowych parametrów wykrywanych impulsów. W tym zakresie zaproponowano rozwiązanie wykorzystujące sieć neuronową do oszacowania początku, końca i maksymalnej amplitudy impulsu o znanym kształcie w sygnale zawierającym mocny szum i trend.

Opracowane modele sieci neuronowych przetestowano pod kątem wpływu kształtu impulsu, mocy zakłóceń szumowych oraz trendu obecnego w sygnałach wejściowych na skuteczność ich działania [1].

2. OPIS ZAPROPONOWANYCH SIECI

2.1. Architektury sieci

Do detekcji impulsów w obecności szumu i trendu opracowano sieć, której strukturę przedstawia rys. 1. Złożona jest ona z 3 warstw typu Linear. Dla pierwszych z nich zastosowano funkcje aktywacji ReLU, a dla ostatniej Sigmoid.

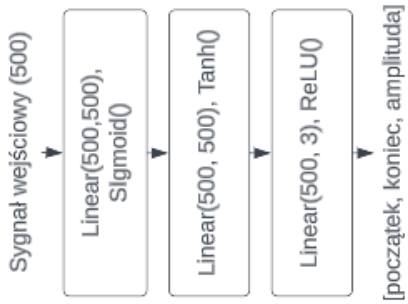


Rys. 1. Architektura sieci klasyfikującej sygnały

Zaproponowana sieć realizuje klasyfikację sygnału, w oparciu o podawane na wejście fragmenty sygnału o długości 500 próbek. Na jej wyjściu otrzymujemy wynik w przedziale <0, 1>, gdzie wartość 0 oznacza pewność, że impuls w sygnale nie występuje, a wartość 1 oznacza pewność sieci, że impuls w sygnale się znajduje.

Strukturę drugiego wariantu sieci neuronowej, służącej do detekcji położenia impulsów w obecności szumu i trendu przedstawiono na rys. 2. Sieć ta złożona jest z 3

warstw typu Linear, dla których użyto kolejno funkcji aktywacji Sigmoid, Tanh, oraz ReLU.



Rys. 2. Architektura sieci szacującej parametry impulsów

Ten wariant sieci również przetwarza sygnały wejściowe o długości 500 próbek, ale na jej wyjściu otrzymujemy trzy parametry liczbowe. Są to kolejno predykcje sieci dotyczące położenia początku impulsu, końca impulsu oraz wartości szczytowej amplitudy impulsu.

W przypadku przetwarzania przez sieć sygnału niezawierającego szukanego impulsu, wystarczającym wskaźnikiem braku impulsu w przetwarzanym sygnale jest wynik, w którym oszacowana wartość szczytowej amplitudy impulsu jest równa zero, zgodnie z logiką, z którą impuls o zerowej amplitudzie odpowiada jego brakowi.

2.2. Trening sieci

Trening zaproponowanych wariantów sieci zrealizowano z wykorzystaniem uczenia nadzorowanego. Zrealizowano go w oparciu o syntetyczne sygnały treningowe wygenerowane wraz z etykietami za pomocą narzędzi opracowanych w [1]. W zależności od badanego czynnika, generowano sygnały wraz z etykietami, zawierające impuls prostokątny lub gaussowski, zawierające trend lub nie oraz zanieczyszczone szumem białym o zadanej mocy. Szczegóły użytych sygnałów dla konkretnego modelu zostały podane przy omówieniu odpowiednich wyników testowania w kolejnych punktach.

W trakcie treningu wariantu pierwszego sieci jako kryterium uczenia zastosowano binarną entropię krzyżową, stanowiącą typową i wydajną funkcję strat dla klasyfikacji binarnej. W drugim wariancie zastosowano kryterium L1Loss. Dla obu opisywanych sieci użyto optymalizatora Adam z współczynnikiem uczenia równym 0,001, czyli standardową wartością.

2.3. Metoda testowania sieci

Efekty działania treningu wypracowanych modeli sieci zostały zweryfikowane na wygenerowanym zbiorze syntetycznych sygnałów testowych i ich etykiet. Do oceny jakości pierwszego wariantu sieci, czyli klasyfikatora binarnego, zastosowano znane miary jakości stosowane w przypadku oceny tego typu klasyfikatorów.

W przypadku drugiego wariantu, do oceny jakości detekcji parametrów impulsu wytrenowanych modeli w procesie walidacji, od każdej wartości w wektorze wynikowym odjęto odpowiadającą jej wartość z wektora wartości oczekiwanych, otrzymując w ten sposób wektor różnic wyników i etykiet. Następnie dla każdej wartości obliczano średnie odchylenie bezwzględne, otrzymując średni błąd estymacji początku, końca i wartości szczytowej amplitudy impulsu. Wyniki otrzymane opisaną metodą omówiono szczegółowo w kolejnych punktach.

Wytrenowane modele sieci neuronowych zostały również zweryfikowane metodą k-krotnej walidacji krzyżowej w celu ewaluacji poprawności procesu uczenia się sieci.

3. WYNIKI WARIANTU 1

3.1. Model wytrenowany za pomocą impulsów prostokątnych w obecności szumu

W celu weryfikacji skuteczności detekcji, sieć w wariancie pierwszym wytrenowano za pomocą wygenerowanych impulsów prostokątnych, zawartych w sygnałach o długości 500 próbek, o losowej wartości amplitudy z przedziału <1, 10>. Do sygnałów dodawany był szum o mocy 10, 15 i 20 dB. Ilość danych pozytywnych (zawierających impuls) i negatywnych (niezawierających impulsu) była jednakowa.

Wynik testowania wytrenowanych modeli za pomocą danych zawierających sygnały o takich samych zakresach wartości amplitudy jak dane testowe, lecz inną wartość mocy szumu, przedstawiono w tab. 1. Wartość precyzji wynosiła w większości przypadków walidacyjnych ponad 90%. Wyjątkiem była sytuacja, gdy moc szumu w sygnałach walidacyjnych była znacznie większa niż wartości mocy szumu w danych treningowych.

Tab. 1. Jakość detekcji impulsów prostokątnych dla różnych mocy szumu w sygnałach testowych

Moc szumu danych treningowych [dB]	Moc szumu danych testowych [dB]	Dokładność	Czułość	Precyzja	NPV
10	-5	0,86	0,99	0,78	0,99
	0	0,96	0,99	0,92	0,99
	5	0,97	0,97	0,98	0,97
	10	0,82	0,65	0,99	0,74
	20	0,51	0,214	0,93	0,52
15	-5	0,5	1	0,5	1
	0	0,64	0,99	0,58	0,99
	5	0,84	0,99	0,76	0,99
	10	0,92	0,93	0,91	0,93
	20	0,53	0,83	0,92	0,51
20	-5	0,5	1	0,5	0
	0	0,5	1	0,5	0
	5	0,61	0,99	0,56	0,99
	10	0,81	0,98	0,73	0,98
	20	0,6	0,23	0,9	0,55

Dokładność dla wszystkich modeli była najlepsza dla mocy szumu danych testowych równej lub o 5-10 dB niższej od mocy szumu w danych treningowych. Wraz ze zwiększaniem poziomu mocy szumu danych testowych sieć miała tendencję do klasyfikacji szumu w sygnałach negatywnych do kategorii pozytywnej. Wyniki dokładności i NPV na poziomie niemalże 50% odpowiadają sytuacji, gdy prawie wszystkie sygnały pozytywne zostały zakwalifikowane fałszywie.

3.2. Model wytrenowany za pomocą impulsów gaussowskich w obecności szumu

Do weryfikacji efektywności sieci w wariancie 1 dla impulsów gaussowskim użyto sygnałów zawierających impulsy gaussowskie o wartościach szczytowych amplitudy z przedziału <1, 10>, losowych wartościach wariancji z przedział <0,05, 0,15>. Do generowanych sygnałów dodawano szum o mocy 5, 10 lub 30 dB. Stosunek ilości danych zawierających impuls do niezawierających był równy jeden.

Walidację skuteczności sieci przeprowadzono w dokładnie taki sam sposób jak w przypadku modelu dla impulsów prostokątnych, z tą różnicą, że zastosowano impulsy gaussowskie o takich samych parametrach jak w danych treningowych. Wyniki testowania sieci dla poszczególnych mocy szumów przedstawiono w tab. 2.

Tab. 2. Jakość detekcji impulsów gaussowskich dla różnych mocy szumu w sygnałach testowych

Moc szumu danych treningowych [dB]	Moc szumu danych testowych [dB]	Dokładność	Czułość	Precyzja	NPV
5	-10	0,8	0,99	0,72	0,99
	0	0,98	0,98	0,98	0,97
	5	0,96	0,94	0,98	0,94
	10	0,9	0,84	0,95	0,86
	20	0,67	0,49	0,76	0,62
	30	0,54	0,33	0,58	0,53
10	-10	0,5	1	0,5	1
	0	0,94	0,98	0,9	0,98
	10	0,91	0,86	0,96	0,87
	20	0,68	0,57	0,74	0,65
	30	0,54	0,61	0,53	0,54
	50	0,5	0,77	0,5	0,5
30	-10	0,5	1	0,5	0
	0	0,5	1	0,5	0
	10	0,5	1	0,5	1
	20	0,6	0,83	0,57	0,69
	30	0,53	0,52	0,53	0,54
	50	0,5	0,67	0,5	0,5

Z poszczególnych wskaźników można wywnioskować, że sieć wyuczona danymi zaszumionymi szumem o danej mocy, radzi sobie równie dobrze z wykrywaniem impulsów w sygnałach słabiej zaszumionych. Wyniki wskaźników dla mocy szumu większej niż wartość treningowa, wskazywały, że sieć większość sygnałów kwalifikuje jako negatywne, zmniejszając przez to wskaźniki NPV i precyzję. Wartości NPV równe 0 widoczne w tabeli wskazują, że sieć zakwalifikowała wszystkie sygnały do kategorii pozytywnej, przez co ilość danych TN również wyniosła 0.

3.3. Model wytrenowany za pomocą impulsów w obecności szumu i trendu liniowego

W celu sprawdzenia wpływu trendu liniowego na skuteczność detekcji sieci w wariancie pierwszym, wytrenowano i przetestowano sieć za pomocą danych o takim samym charakterze jak w p. 3.2, lecz z dodanym trendem liniowym.

Tab. 3. Wyniki testowania sieci wytrenowanej za pomocą impulsów gaussowskich w obecności szumu i trendu

Moc szumu danych treningowych [dB]	Moc szumu danych testowych [dB]	Dokładność	Czułość	Precyzja	NPV
5	-5	0,93	0,86	0,99	0,88
	0	0,92	0,85	0,99	0,87
	5	0,9	0,85	0,95	0,86
	10	0,81	0,84	0,8	0,83
	15	0,69	0,8	0,65	0,74
10	-5	0,87	0,75	1	0,8
	0	0,87	0,75	0,99	0,8
	5	0,87	0,76	0,99	0,8
	10	0,84	0,77	0,9	0,8
	15	0,7	0,81	0,66	0,76
	20	0,55	0,87	0,53	0,66
15	-5	0,81	0,63	1	0,73
	0	0,81	0,63	1	0,73
	5	0,82	0,65	0,99	0,74
	10	0,81	0,65	0,96	0,74
	15	0,73	0,68	0,76	0,71
	20	0,6	0,78	0,58	0,65

Wyniki testowania sieci w zależności od mocy szumu w danych testowych przedstawiono w tab. 3. Dodanie trendu do sygnałów sprawiło, że dokładność predykcji dla danych walidacyjnych o tej samej lub mniejszej mocy szumu niż dla danych treningowych zmniejszyła się o około 0,07, lecz dla wartości mocy szumu większych było odwrotnie i była ona większa. Wartość precyzji dla mocy szumu takiej samej lub mniejszej jak w danych treningowych była zbliżona. Była ona jednak dużo niższa dla większej mocy szumu. Wskazania wskaźnika NPV były wyższe, a wskaźnika czułości niższe.

Dodanie trendu liniowego do danych pogorszyło jakość predykcji wytrenowanych modeli. Sieć osiągała nieco gorszą jakość predykcji niż dla sygnałów bez trendu liniowego, przy tych samych wartościach mocy szumu.

4. WYNIKI WARIANTU 2

4.1. Model wytrenowany za pomocą impulsów prostokątnych

W celu weryfikacji skuteczności detekcji sieci w wariancie drugim, sieć wytrenowano za pomocą wygenerowanych impulsów prostokątnych, o takich samych parametrach jak w p. 3.1, z tą różnicą, że moc szumu była wartością losową z przedziału <0, 30> dB, użyto również impulsów niezawierających szumu.

Zastosowanie do treningu sygnałów niezanieczyszczonych szumem pozwoliło na zwiększenie precyzji sieci dla sygnałów zanieczyszczonych szumem z dolnej wartości przedziału mocy.

Wyniki testowania w zależności od mocy szumu danych testowych przedstawiono w tab. 4. Charakter danych testowych był taki sam jak treningowych z tą różnicą, że moc szumu dla poszczególnych modeli sieci była stała i przyjmowała wartości od -5 do 25 dB z krokiem 5 dB.

Tab. 4. Wyniki walidacji drugiej sieci wytrenowanej za pomocą impulsów prostokątnych

moc szumu [dB]	-5	0	5	10	15	20	25
średnia bł. początku	0	0	-1	-3	-8	-20	-40
odch. st.. bł. początku	7,36	10,33	15,26	26,41	42,06	61,25	80,9
średnia bł. końca	-1	-1	-2	-4	-15	-44	-102
odch. st. bł. końca	7,66	10,77	17,23	34,14	67,27	113,85	157,36
średnia bł. amplitudy	1,33	1,33	1,23	1,18	0,79	0,03	-1,29
odch. st.. bł. amplitudy	2,59	2,59	2,58	2,51	2,33	2,17	2,17
odsetek TN [%]	99,61	99,56	99,49	98,96	98,4	98,07	97,67

Sieć w zadowalający sposób radziła sobie z estymacją wartości wektora wyjściowego dla sygnałów niezawierających impulsów, prawidłowo podając wynik w około 99% przypadków dla mocy szumów od -5 do 10 dB. Dla większych mocy szumu odsetek ten spada o około 0,4% na każde dodatkowe 5 dB mocy szumu.

4.2. Model wytrenowany za pomocą impulsów gaussowskich

Do weryfikacji efektywności sieci w wariancie 1 dla impulsów gaussowskim użyto sygnałów o takich samych parametrach, jak w p. 3.2 i losowej wartości mocy szumu z przedziału od -5 do 20 dB.

Walidację skuteczności sieci, której wyniki przedstawiono w tab. 5, przeprowadzono w dokładnie taki sam sposób jak w przypadku modelu dla impulsów prostokątnych, z tą różnicą, że zastosowano impulsy gaussowskie o takich samych parametrach jak w danych treningowych.

Klasyfikacja sygnałów niezawierających impulsów pogorszyła się w stosunku do modelu dla impulsów prostokątnych i wyniosła od około 80% do 88%. Zaobserwowano, że dla klasyfikacji tych sygnałów wraz ze wzrostem mocy szumu pogorszyły się wskaźniki dla sygnałów zawierających impulsy.

Tab. 5. Wyniki walidacji drugiej sieci wytrenowanej za pomocą impulsów gaussowskich

moc szumu [dB]	-10	-5	0	5	10	15	20
średnia bł. początku	-4	-4	-6	-12	-28	-62	-110
odch. st.. bł. początku	20,53	25,98	33,87	57,17	87,83	121,2	146,4
średnia bł. końca	2	1	0	-8	-33	-88	-174
odch. st.. bł. końca	21,22	28,25	37,08	65,68	11,09	163,6	193
śr. bł. amplitudy	0,03	-0,02	-0,11	-0,32	-0,83	-1,74	-3,02
odch. st.. bł. amplitudy	2,57	2,54	2,55	2,36	2,12	1,92	2,05
odsetek TN [%]	79,69	79,64	79,34	79,79	81,25	83,62	88,35

Średnie błędów estymacji pozycji impulsów i ich odchylenia standardowe rosły wraz ze wzrostem mocy szumu w danych walidacyjnych, przy czym odchylenia standardowe miały większe wartości w stosunku do wyników dla impulsów prostokątnych.

Wartości średnie błędu estymacji wartości amplitudy dla mocy szumu od -10 do 5 dB dawały wyniki na poziomie od -0,32 dla 5 dB do 0,03 dla -10 dB, co oznacza, że dla tych wartości sieć dosyć precyzyjnie wskazywała wyniki.

4.3. Testowanie za pomocą impulsów prostokątnych w obecności szumu i trendu.

Model sieci został wytrenowany sygnałami o takim samym charakterze jak w punkcie nieuwzględniającym trendu, lecz z dodatkowym trendem liniowym.

Wyniki walidacji modelu przedstawiono w tab. 6. Najbardziej zauważalną różnicę, w stosunku do sieci nieuwzględniającej trendu, jest znaczący spadek liczby prawidłowej predykcji dla sygnałów ich niezawierających. Dla większości mocy szumu, wynik stanowiły wartości mniejsze niż 50%, co jest znacznie gorszym wynikiem niż dla sieci nieuwzględniającej trendu.

Średnie wartości błędów estymacji początku i końca impulsów są niższe w porównaniu dla modeli bez trendu, a ich odchylenie standardowe większe o około 50% dla mniejszych wartości mocy szumów (od -5 dB do 5 dB). Dla wszystkich wartości mocy szumu w danych testowych sieć estymuje wartości amplitudy z niedomiarem, a odchylenie standardowe tych wartości rośnie wraz ze wzrostem mocy zakłóceń w sygnale.

Tab. 6. Wyniki walidacji sieci wytrenowanej za pomocą impulsów prostokątnych w obecności szumu i trendu

moc szumu [dB]	-5	0	5	10	15	20	25
średnia bł. początku	-3	-3	-3	-5	-7	-11	-22
odch. st.. bł. początku	19,47	19,63	22,87	29,77	39,3	55,12	74,01
średnia bł. końca	-2	-2	-1	-2	-5	-16	-58
odch. st.. bł. końca	15,59	17,95	21,96	31,32	50,35	79,1	130,1
śr. bł. amplitudy	-0,19	-0,22	-0,22	-0,3	-0,46	-0,61	-0,97
odch. st.. bł. amplitudy	1,47	1,49	1,55	1,72	1,96	2,3	2,68
odsetek TN [%]	15,75	15,58	17,73	21,04	27,54	41,18	59,17

4.4. Testowanie za pomocą impulsów gaussowskich w obecności szumu i trendu.

Trening sieci w wariancie 2 za pomocą impulsów z p. 4.2, lecz z dodanym trendem liniowym okazał się nieefektywny. Wykres strat na przestrzeni treningu wskazywał na brak efektów uczenia się sieci, przez co wywnioskowano, że dodanie trendu do danych uniemożliwia sieci o zadanej architekturze naukę poprawnej detekcji impulsów w sygnałach o takim charakterze.

4.5. Testowanie wariantów za pomocą 5-krotnej walidacji krzyżowej

W celu sprawdzenia poprawności wytrenowanych modeli, przeprowadzono dla nich 5-krotną walidację krzyżową. Uzyskane wyniki dla danych treningowych i walidacyjnych zbiegały do tej samej wartości na przestrzeni epok treningu, co wskazuje na brak przetrenowania i przecieków danych.

5. PODSUMOWANIE

W pracy przedstawiono dwie sieci neuronowe, które udało się zastosować do detekcji występowania impulsów oraz szacowania ich położenia oraz amplitudy szczytowej. Wyniki przeprowadzonych eksperymentów wskazują na to, że zmniejszanie SNR sygnałów niekorzystnie wpływało na detekcję impulsów w obu opracowanych wariantach sieci, od pewnego poziomu ją uniemożliwiając. Dodanie trendu liniowego obniżyło graniczną moc zakłóceń, przy których wyniki działania opracowanych modeli były zadowalające.

Wyniki uzyskane dla neuronowego detektora położenia i estymacji amplitudy impulsów jasno sugerują wpływ charakteru zboczy impulsów na skuteczność detekcji, wskazując na korzystny wpływ bardziej stromych zboczy na detekcję.

W przyszłych rozważaniach planowane jest przetestowanie innych, bardziej złożonych, architektur sieci neuronowych. Być może większa złożoność sieci pozwoli na zmniejszenie granicznej wartości SNR sygnałów, dla których detekcja jest skuteczna.

LITERATURA

[1] Betka Szymon 2022, Zastosowanie sieci neuronowych do detekcji impulsów o znanym kształcie w obecności silnego szumu i trendu, projekt dyplomowy inżynierski, WETI PG

[2] Hu J.-F. i in. 2015, “Harmonic signal detection method from strong chaotic background based on optimal filter” Acta Physica Sinica, vol. 64, no. 22, 220504.

[3] Ke D. i in. 2019, “Blind detection techniques for non-cooperative communication signals based on deep learning,” IEEE Access, vol. 7, pp. 89218–89225.

[4] Liyun Su i in. 2019, "Detection and Extraction of Weak Pulse Signals in Chaotic Noise with PTAR and DLTAR Models", Mathematical Problems in Engineering, vol. 2019, 4842102.

[5] Mendis G. J. i in. 2019, “Deep learning based radio-signal identification with hardware design,” IEEE Transactions on Aerospace and Electronic Systems, vol. 55, no. 5, pp. 2516–2531.

[6] Xing H.-Y. i in. 2012, “Detection of weak target signal with least-squares support vector machine and generalized embedding windows under chaotic background,” Acta Physica Sinica, vol. 61, no. 10, 100506.

[7] Yuan Y. i in. 2019, “DeepMorse: a deep convolutional learning method for blind Morse signal detection in wideband wireless spectrum,” IEEE Access, vol. 7, pp. 8057 80587.

[8] Yue L. i in. 2006, “Weak effective seismic signal detection capability of chaotic oscillator detection system,” Chinese Science Bulletin, vol. 51, no. 14, pp. 1710–1716

SYSTEM ŁĄCZNOŚCI DLA SATELITÓW CUBESAT – ARCHITEKTURA I PROTOKOŁY

CUBESAT SATELLITES COMMUNICATION SYSTEM – ARCHITECTURE AND PROTOCOLS

Rafał Krenz[1]; Marek Michalski[2]; Ilia Zainutdinov[3]

[1] Politechnika Poznańska, Poznań, rafal.krenz@put.poznan.pl
[2] Politechnika Poznańska, Poznań, marek.michalski@put.poznan.pl
[3] Politechnika Poznańska, Poznań, ilia.zainutdinov@doctorate.put.poznan.pl

DOI: 10.15199/59.2023.4.48

Streszczenie: **Konstruowane w ostatnich latach satelity standardu Cubesat zawierają coraz bardziej rozbudowane przyrządy badawcze, będące źródłem znacznej ilości informacji. Przesłanie jej na ziemię wymaga nowych, bardziej efektywnych systemów transmisyjnych. W referacie zaprezentowano architekturę systemu skonstruowanego na Politechnice Poznańskiej oraz opisano protokoły transmisyjne umożliwiające uzyskanie dużej sprawności transmisji. Znajdzie on zastosowanie w dwóch misjach przewidzianych na lata 2023-24.**

Abstract: **The complexity of the payload installed onboard Cubesat satellites is still growing. The devices produce increasing amounts of information. Therefore new and more efficient communication systems for downloading satellite data must be developed. In this paper an architecture and communication protocols of a system developed and implemented in Poznań University of Technology is presented. It will be used during two missions planned for 2023-24.**

Słowa kluczowe: **satelity Cubesat, łącza satelitarne, adaptacja łącza, uczenie maszynowe, automatyczna retransmisja**

Keywords: **Cubesat satellite, satellite link, link adaptation. machine learning, Automatic Repeat Request**

1. WSTĘP

Wraz z rosnącą popularnością standardu Cubesat konstruowane są coraz większe jednostki o rozmiarach 3U-6U-12U. Umieszczane na nich przyrządy badawcze (kamery pracujące w różnych zakresach widma, radary SAR, itp.) są źródłem znacznej ilości informacji, która musi zostać przesłana do stacji naziemnej w celu jej obróbki. Rosną więc również wymagania stawiane systemom transmisyjnym, które powinny pozwalać na uzyskanie dużej sprawności transmisji przy ograniczeniach wynikających z konstrukcji satelitów Cubesat, w szczególności energetycznych oraz fizycznych rozmiarów modułów komunikacyjnych.

Współpraca nawiązana przez Politechnikę Poznańską z Politechniką Kijowską oraz firmą SatRev[1] zaowocowała skonstruowaniem systemu transmisji danych o nazwie CBSR[2] przeznaczonego dla standardu Cubesat, pracującego w pasmie C w kanale o szerokości do 20MHz. Zaprojektowany i zbudowany system jest całkowicie oryginalny i wykorzystuje wyłącznie własne rozwiązania. co czyni go niekompatybilnym z dostępnymi na rynku systemami, zapewnia jednak jego znaczną elastyczność i uniwersalność.

Ponieważ warstwa fizyczna systemu została zaprezentowana już wcześniej [1], w dalszej części artykułu skupiono się na przedstawieniu jego architektury oraz protokołów komunikacyjnych. W rozdziale 2 opisano architekturę systemu, ze szczególnym uwzględnieniem segmentu naziemnego. Rozdział 3 poświęcony jest protokołom komunikacyjnym, których zadaniem jest zapewnienie wiarygodnej i efektywnej transmisji danych z satelity do stacji naziemnej. Za zarządzanie parametrami transmisji odpowiedzialny jest moduł uczenia maszynowego, opisany w rozdziale 4. Rozdział 5 stanowi podsumowanie.

2. ARCHITEKTURA SYSTEMU CBSR

Realizacja projektu z partnerami będącymi operatorami satelitów Cubesat, na pokładzie których zainstalowane zostaną moduły nadawcze CBSR wymusiła rozproszoną architekturę segmentu naziemnego. W konsekwencji powstał system charakteryzujący się znaczną elastycznością i uniwersalnością, zapewniającą możliwość dalszego wzrostu efektywności jego działania.

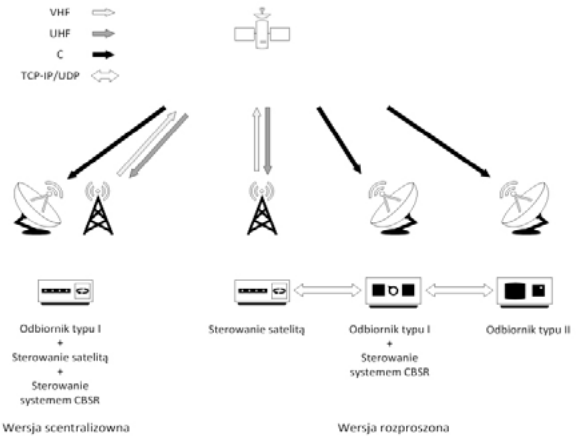
Do odbiory sygnałów danych w pasmie C przewidziano wykorzystanie dwóch klas odbiorników (Rys.1):

- Typ I – odbiornik sprzętowy, zrealizowany z wykorzystaniem układów FPGA, realizujących procedury przetwarzania sygnałów w pasmie podstawowym w czasie rzeczywistym
- Typ II – odbiornik programowy, zrealizowany na platformie GNURadio, realizującej procedury przetwarzania sygnałów w pasmie podstawowym w trybie *off-line*, współpracujący z modułem radiowym SDR, np. USRP B210 firmy Ettus Research

Ponieważ system umożliwia jedynie transmisję jednokierunkową, dla poprawnego funkcjonowania konieczne jest wykorzystanie standardowego łącza telemetryczno-sterującego (VHF/UHF). Są nim przesyłane informacje sterujące, dotyczące retransmisji błędnie

[1] Dalej nazywane odpowiednio PP, KPI oraz SR
[2] C-Band Satellite Radio

zdekodowanych pakietów danych (patrz rozdział 3) oraz tzw. plany sesji, definiujące m.in. parametry transmisji dla kolejnych sesji komunikacyjnych (patrz rozdział 4).



Rys. 1. Architektura segmentu naziemnego CBSR

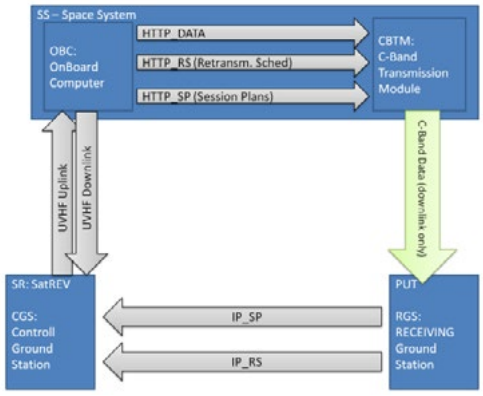
Ponieważ operatorami misji, w których po raz pierwszy wykorzystany zostanie system CBSR są KPI oraz SR, zaś odbiór sygnałów w pasmie C możliwy jest obecnie wyłącznie w PP, informacje sterujące oraz zdekodowane dane użytkowe przesyłane są łączami internetowymi z PP do KPI/SR. Możliwa jest również rozbudowa systemu o dodatkowe stacje odbiorcze wyposażone w odbiorniki typu II, których zadaniem jest wyłącznie przesyłanie łączami internetowymi zdekodowanych pakietów danych do PP. Oprogramowanie realizujące odbiornik typu II będzie dostępne jako otwarto źródłowe, a operatorzy zainteresowani jego wykorzystaniem będą mogli brać aktywny udział w sesjach komunikacyjnych po zarejestrowaniu się w odpowiedniej bazie danych PP. Pozwoli to na zwiększenie efektywności systemu poprzez zwiększenie liczby sesji komunikacyjnych, które w przypadku satelitów LEO trwają zaledwie kilka minut.

Za zarządzanie pracą systemu CSBR, a więc kompletowanie poprawnie zdekodowanych pakietów i odtwarzanie przesyłanych plików, generowanie żądań retransmisji błędnych pakietów oraz przygotowanie planów sesji dla wszystkich zarejestrowanych stacji odbiorczych, odpowiedzialny jest moduł sterowania, zlokalizowany na PP i współpracujący z odbiornikiem typu I.

W przypadku, gdy operator satelity dysponuje pełną infrastrukturą komunikacyjną (pasma VHF/UHF oraz C) wszystkie funkcje systemu CBSR mogą zostać oczywiście scentralizowane. Nie wyklucza to oczywiście rozbudowa systemu o dodatkowe stacje odbiorcze wyposażone w odbiorniki typu II.

3. PROTOKOŁY KOMUNIKACYJNE

Niniejszy rozdział przedstawia opis mechanizmów wykorzystywanych do transmisji danych powyżej warstw fizycznej. W budowie satelity (Space System) można wyróżnić kilka głównych modułów (Rys. 2): OBC – On Board Computer oraz CBTM – C-Band Transmission Module. Zespół PP jest odpowiedzialny za moduł transmisji danych w paśmie C, natomiast zespół SR realizuje moduł nadzorujący pracę satelity oraz dwustronną komunikację telemetryczno-sterującą w pasmach UHF/VHF.



Rys. 2. Moduły architektury systemu

To właśnie transmisja danych z modułu umieszczonego na satelicie do stacji naziemnej jest głównym zadaniem modułu komunikacyjnego. Dane te dostarczone są przez moduł partnera projektowego. Są to dane pozyskane z różnych urządzeń pomiarowych i rejestracyjnych, głównie fotografie wykonane na zlecenie klientów i partnerów projektu. Nasz moduł nie musi być świadomy treści, dla niego są to pliki cyfrowe o unikatowych nazwach. Są one dostępne poprzez specjalnie przygotowane Web-API (http), które jest wykorzystywane do przesyłania danych o dużym wolumenie pomiędzy modułami systemu. Działa ono na bazie protokołu Ethernet i pozwala zunifikować kontakt pomiędzy elementami systemu, ale nie jest to jedyny sposób realizacji takiej komunikacji.

O ile nadawanie odbywa się z jednego satelity, odbiór może być realizowany przez wiele naziemnych stacji odbiorczych. Dzięki temu, że mogą one być odległe geograficznie, są w stanie znacznie wydłużyć czas przebywania satelity w zasięgu całego systemu. Z uwagi na wysokość orbity, satelita nie jest nieruchomy względem punktów na ziemi, porusza się ze znaczną prędkością, co wymusza wykorzystanie złożonych mechanizmów śledzenia i naprowadzania anten odbiorczych oraz skutkuje wyłącznie czasową obecnością satelity w zasięgu poszczególnych naziemnych stacji odbiorczych. Należy zwrócić uwagę, że moduł CBTM pracuje w trybie jednokierunkowym, tzn. satelita nadaje w kierunku ziemi i może nadawać do dowolnej stacji odbiorczej, natomiast dodatkowe łącze w górę jest realizowane przez specjalny moduł, obsługiwane jest ono tylko przez wybrane, w pełni nadzorowane stacje nadawczo-odbiorcze (w skrajności liczba takich stacji jest zredukowana do jednej). Oznacza to, że odbiór może być realizowany w dużo szerszym zakresie niż nadawanie do satelity, co powoduje wymóg precyzyjnego planowania działania systemu. Jest to realizowane poprzez mechanizm sesji transmisyjnych. System jest w pełni przewidywalny, jest możliwość poznania dokładnej aktualnej pozycji satelity oraz estymacji jego przyszłych pozycji, czyli również momentów czasu, w których będzie możliwy odbiór przez poszczególne stacje odbiorcze. Te dane, wraz z

wnioskami z monitorowania jakości dotychczasowych transmisji, spodziewanymi warunkami atmosferycznymi, pozwalają efektywnie zaplanować poszczególne sesje transmisyjne. Każda sesja ma inne parametry, są one obliczane indywidualnie i przesyłane odpowiednio wcześniej do satelity jako tzw. plany sesji w celu zaprogramowania aktywności jego poszczególnych modułów. Dzięki temu, na czas nieaktywności, moduł transmisyjny jest wyłączany, a włączany jest chwile przed wejściem w zasięg stacji odbiorczej, tylko na czas realizacji transmisji, co znacznie poprawia bilans energetyczny całego systemu. Co więcej, zakładana prędkość transmisji danych zmienia się adekwatnie do geometrycznej odległości od wybranej stacji odbiorczej. Takie podejście pozwala dopasować parametry transmisji w celu zmaksymalizowania efektywności wykorzystania łącza w dół. Podczas poszczególnych sesji transmisyjnych, odpowiednio fragmentowane i opatrzone identyfikatorami dane są przesyłane na ziemię, gdzie są odtwarzane pierwotne pliki danych. Moduł na satelicie pracuje samodzielnie, na bieżąco monitoruje postęp transmisji (liczbę wysłanych danych, liczbę plików). Po upływie sesji transmisyjnej – stan systemu jest zapisany, aby w kolejnej sesji kontynuować transmisję od precyzyjnie wyliczonego punktu, a nie od początku strumienia danych. W stacji naziemnej, po zebraniu odebranych danych, jest generowana informacja o koniecznych retransmisjach niepoprawnie odebranych fragmentów oraz o potwierdzeniu poprawnie odebranych całych plików. Komunikaty te trafiają do satelity w momencie, gdy możliwe jest wykorzystanie łącza w górę, wówczas następuje zaplanowanie retransmisji danych zidentyfikowanych jako błędnie odebrane oraz usunięcie danych, których poprawną transmisję potwierdzono oraz aktualizacja informacji o stanie sytemu. Co bardzo przydatne, dane odebrane mogą być dostarczane do centrali na bieżąco z odbiorników typu I, oraz w trybie z opóźnieniem, w postaci pliku zawierającego dane zebrane podczas całej sesji komunikacyjnej przez odbiorniki typu II. Opracowane oprogramowanie narzędziowe dokonuje konkatenacji danych do jednego spójnego strumienia danych odebranych przez system, bez rozróżnienia, który typ stacji odbiorczej dostarczył poszczególne fragmenty. Zbierane są odpowiednie dane statystyczne pozwalające ocenić pracę systemu oraz estymować parametry przyszłych sesji komunikacyjnych. Sposób i intensywność generowania potwierdzeń do satelity są adaptacyjnie ustalane w zależności od jakości transmisji, ilości danych w potwierdzeniach oraz stanu systemu. Potwierdzenia te są przesyłane od centrali do stacji dysponującej łączem telemetryczno-sterującym, a następnie do satelity w momencie, gdy jest on w jej zasięgu.

4. STEROWANIE PARAMETRAMI TRANSMISJI – UCZENIE MASZYNOWE

Uczenie maszynowe jest dziedziną sztucznej inteligencji, która umożliwia systemom komputerowym automatyczne uczenie się i poprawianie swojego działania na podstawie doświadczenia, bez potrzeby jawnego programowania. Istnieją różne rodzaje uczenia maszynowego, w tym uczenie nadzorowane, uczenie nienadzorowane i uczenie ze wzmocnieniem. Każdy z tych rodzajów ma swoje unikalne zalety i wady, a wybór odpowiedniego algorytmu uczenia maszynowego jest kluczowy dla sukcesu każdego projektu [2,3]. W opisanym systemie technika ta zostanie wykorzystana do optymalizacji parametrów transmisji, mającej na celu maksymalizację wolumenu przesyłanych z satelity danych.

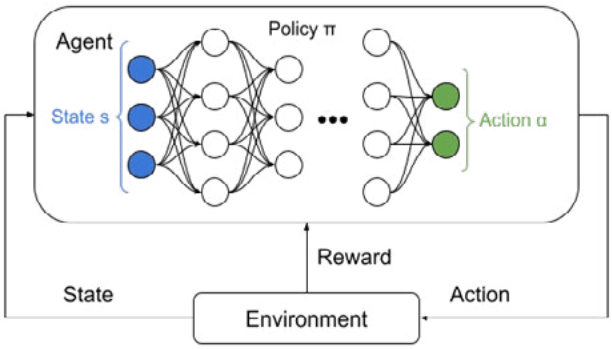
Uczenie nadzorowane to rodzaj uczenia maszynowego, w którym algorytm uczy się przewidywania wartości wyjściowej na podstawie danych wejściowych i odpowiadających im oznaczonych przykładów. Innymi słowy, algorytm jest szkolony na zbiorze danych, gdzie pożądany wynik jest już znany. Następnie algorytm uczy się mapowania danych wejściowych na wyjście poprzez minimalizowanie różnicy między przewidywanymi a rzeczywistymi wartościami wyjściowymi. Tego typu uczenie maszynowe jest powszechnie stosowane w problemach klasyfikacji i regresji, takich jak rozpoznawanie obrazów czy przewidywanie cen akcji. W kontekście komunikacji satelitarnej może być trudno lub niemożliwe uzyskanie oznaczonych danych treningowych do nauki. Wynika to z tego, że często trudno jest symulować lub przewidzieć dokładne warunki, jakie satelita napotka podczas rzeczywistych sesji komunikacyjnych. Ponadto charakter danych w komunikacji satelitarnej może być silnie zmienny, co utrudnia stworzenie wyczerpującego oznakowanego zbioru danych. Nawet gdyby można było uzyskać oznakowany zbiór danych, może on nie być reprezentatywny dla wszystkich możliwych scenariuszy, a model może nie uogólniać dobrze do nowych, nieznanych sytuacji.

Uczenie nienadzorowane to rodzaj uczenia maszynowego, w którym algorytm uczy się wykrywać wzorce lub strukturę w danych nieoznakowanych, bez wyraźnego poinformowania, czego ma szukać. Algorytm jest szkolony na zbiorze danych bez żadnych oznakowanych przykładów i uczy się identyfikować podobieństwa, różnice i grupy w danych. Ten rodzaj uczenia maszynowego jest powszechnie stosowany do grupowania, wykrywania anomalii i redukcji wymiarowości. W komunikacji satelitarnej dane są bardzo zorganizowane, a cel jest jasny - optymalizacja wydajności komunikacji. Algorytmy uczenia nienadzorowanego mogą nie być odpowiednie do tego zadania, ponieważ nie mają konkretnego celu do optymalizacji i mogą nie być w stanie podejmować decyzji prowadzących do poprawy wydajności. Ponadto, algorytmy uczenia nienadzorowanego mogą wymagać dużej ilości danych, aby skutecznie odkrywać wzorce, co może być trudne do uzyskania w systemach komunikacji satelitarnej, gdzie często występują ograniczone możliwości zbierania danych. Dlatego uczenie nienadzorowane może nie być najbardziej efektywnym

ani skutecznym podejściem do optymalizacji wydajności komunikacyjnej w systemach satelitarnych.

Uczenie ze wzmocnieniem wyróżnia się swoją zdolnością do nauki poprzez interakcję z otoczeniem za pomocą prób i błędów. Ten podejście zostało skutecznie zastosowane w różnych dziedzinach, takich jak robotyka, gry i systemy autonomiczne [4]. Algorytmy uczenia ze wzmocnieniem (ang. Reinforced Learning - RL) mogą być stosowane do optymalizacji sterowania systemu komunikacyjnego między stacją naziemną a satelitą. W tym konkretnym zastosowaniu algorytm RL ma za zadanie sterowanie dwoma parametrami: szerokością pasma transmisji, które jest wybierane raz dla całego okna transmisji, oraz sprawnością kodowania, która może być zmieniany wielokrotnie podczas sesji. Ważne jest zauważenie, że nie wszystkie sprawności kodowania są stosowane dla większych pasm. Dlatego algorytm RL musi brać pod uwagę tę relację między pasmem a współczynnikiem kodowania przy wyborze optymalnych wartości.

Zaprezentowany na Rys. 3 schemat blokowy algorytmu uczenia ze wzmocnieniem ułatwi zrozumienie wykorzystania tej metody do optymalizacji wolumenu danych w systemie łączności satelitarnej.



Rys. 3 Schemat blokowy na wysokim poziomie dla algorytmu uczenia ze wzmocnieniem (RL)

State: Stan to aktualne środowisko informacyjne, które jest przetwarzane przez algorytm RL w celu podjęcia decyzji. W tym przypadku stan obejmuje parametry, takie jak elewacja satelity, odległość do satelity, straty atmosferyczne itp., które są obliczane na podstawie danych z elementu dwuliniowego (TLE) i innych źródeł. Aby zoptymalizować wydajność komunikacji, algorytm RL musi przetwarzać całą trajektorię ścieżki satelity.

Reward: Nagroda to wartość skalarna, którą algorytm RL otrzymuje po każdej akcji i służy do oceny, jak dobrze działa agent. W tym przypadku nagroda opiera się na wydajności komunikacji, takiej jak całkowita ilość danych, która została pomyślnie przesłana. Celem jest maksymalizacja nagrody w czasie.

Action: Akcja to decyzja, którą algorytm RL podejmuje na podstawie bieżącego stanu. W tym przypadku algorytm RL kontroluje dwa parametry: szerokość pasma, która jest wybierana raz na całe okno transmisji, oraz sprawność kodowania, która może być zmieniany wielokrotnie podczas sesji. Algorytm wybiera optymalną szerokość pasma, która może być używana przez całe okno transmisji i zapewnia znaczniki czasowe, kiedy sprawność kodowania musi być zmieniony podczas okna transmisji. Sprawność kodowania jest dostosowywana na podstawie bieżących warunków kanału, aby maksymalizować efektywną przepustowość przy zachowaniu niezawodnej transmisji.

Policy: Polityka to strategia algorytmu RL w zakresie wyboru działań na podstawie bieżącego stanu. W tym przypadku polityka polega na uczeniu się optymalnej polityki, która mapuje bieżący stan na najlepsze działania, maksymalizując nagrodę w czasie. Algorytm RL jest szkolony na podstawie danych historycznych, aby poprawić swoją wydajność w czasie, umożliwiając systemowi dostosowanie się do zmieniających się warunków i poprawienie ogólnej niezawodności systemu. Analizując przeszłe sesje komunikacyjne, algorytm może nauczyć się identyfikować wzorce i optymalizować parametry kontrolne, aby osiągnąć lepszą wydajność w przyszłych sesjach. Po szkoleniu i wdrożeniu algorytmu RL może działać autonomicznie, dostosowując parametry kontrolne do optymalizacji wydajności komunikacji bez potrzeby ciągłego nadzoru ze strony człowieka.

5. PODSUMOWANIE

Zapewnienie efektywnej komunikacji z satelitami Cubesat, realizowanej przy użyciu stosunkowo nieskomplikowanych i łatwo dostępnych urządzeń jest znaczącym wyzwaniem. Sądzimy, że zaproponowany przez nas system może przyczynić się do szerszego wykorzystania nanosatelitów do przeprowadzania wszelkiego rodzaju eksperymentów naukowych organizowanych przez ośrodki badawcze i akademickie a nawet stowarzyszenia np. radioamatorów. Praktyczne potwierdzenie jego efektywnego i niezawodnego działania możliwe będzie dzięki zaplanowanym do realizacji w latach 2023-24 misjom kosmicznym. Ich wynikami podzielimy się na pewno podczas kolejnych edycji konferencji KRiT!

Projekt finansowany w ramach grantu NCBR POIR.01.01.01-00-1211/19-00 oraz subwencji MNiSW 0312/SBAD/8162.

LITERATURA

[1] Rafał Krenz, Michał Sybis, Ilia Zainutdinov, *Wykorzystanie pasma C do łączności z nano-satelitami CUBESAT*, Przegląd Telekomunikacyjny - Wiadomości Telekomunikacyjne - 2022, nr 4, s. 371-375.

[2] Goodfellow Ian, Bengio Yoshua, Courville Aaron. 2016. "Deep Learning". Książka

[3] Géron Aurélien. 2019. "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems". Książka: 806-809

[4] S. Sutton Richard, G. Barto Andrew. 2018. "Reinforcement Learning: An Introduction". Książka: 53-59

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

FireFinder - Sieć detektorów środowiskowych, wykorzystujących AI i NB-IoT umożliwiających wczesną detekcję pożarów w lasach

FireFinder - A network of environmental detectors using AI and NB-IoT for early detection of forest fires

Konrad Zuchniak[1]; Eryk Kosk[2]; Piotr Jaglarz[3];

[1] Neuralbit Technologies, Kraków, konrad.zuchniak@neuralbit.pl
[2] Neuralbit Technologies, Kraków, eryk.kosk@neuralbit.pl
[3] Neuralbit Technologies, Kraków, piotr.jaglarz@neuralbit.pl

DOI: 10.15199/59.2023.4.49

Streszczenie: FireFinder to system do wczesnego wykrywania pożarów w lasach. Czujniki FireFinder stale monitorują parametry powietrza w poszukiwaniu anomalii. W przypadku wykrycia nawet śladowych ilości dymu, pojedynczy czujnik przesyła informacje do centralnego serwera oraz wywołuje okoliczne czujniki do częstszego przesyłania pomiarów. Algorytmy uczenia maszynowego analizują w czasie rzeczywistym dane otrzymane z sieci czujników w celu wyeliminowania fałszywych alertów oraz ustalenia dokładnej lokalizacji pożaru i kierunku jego rozwoju. W przypadku potwierdzenia zagrożenia odpowiednie służby lokalnej straży pożarnej są alarmowane automatycznie. System przesyła informacje o lokalizacji pożaru, jego przewidywanej wielkości oraz dodatkowe prognozy dotyczące kierunku i szybkości rozprzestrzeniania się ognia. Niniejsza publikacja przedstawia opis projektu oraz wybrane zagadnienia inżynieryjno-naukowe, które stanowiły trzon realizacyjny projektu. Projekt pod tytułem "FireFinder - innowacyjny system do wczesnego wykrywania pożarów w lasach" był współfinansowany ze środków Funduszy Europejskich w ramach Regionalnego Programu Operacyjnego Województwa Małopolskiego 2014-2020.

Abstract: FireFinder is an system for early detection of forest fires. FireFinder sensors constantly monitor air parameters in search of anomalies. If even trace amounts of smoke are detected, a single sensor sends information to the central server and calls nearby sensors to send measurements more often. Machine learning algorithms analyze the data received from the sensor network in real time to eliminate false alerts and determine the exact location and direction of the fire. If a threat is confirmed, the relevant local fire services are alerted automatically. The system sends information about the location of the fire, its expected size and additional forecasts regarding the direction and speed of fire spread. This publication presents a description of the project and selected engineering and scientific issues that formed the core of the project.

Słowa kluczowe: detekcja pożarów, Narrowband IoT, sensory środowiskowe, sieci sensorowe, uczenie maszynowe

Keywords: environmental sensors, fire detection, machine learning, Narrowband IoT, sensor networks

1. WSTĘP

Co roku na świecie wybuchają setki tysięcy pożarów lasów. Szczególnie w ostatnich latach, w związku z ociepleniem klimatu doszło do wielu niekontrolowanych pożarów na olbrzymią skalę. Są one powiązane z utratą życia przez wiele osób i zwierząt. Dodatkowo niosą one ze sobą wielkie straty finansowe (każdy spalony hektar to strata na poziomie 20 tys. zł związana z utratą surowca drzewnego). Celem czujników FireFinder jest umożliwienie szybszego wykrycia pożaru, wcześniejszego rozpoczęcia akcji gaśniczej, a dzięki temu ograniczenie strat. Aktualnie najpopularniejszym rozwiązaniem do wykrywania pożarów lasów są wieże obserwacyjne. Ten system ma jednak pewne wady. Przede wszystkim za ich pośrednictwem możliwe jest wykrycie pożaru dopiero w momencie, kiedy dym będzie widoczny ponad koronami drzew, co znacząco ogranicza zdolność detekcji wczesnych pożarów, zwłaszcza w wietrzne dni. Dodatkowo wieże takie nie spełniają swojej funkcji w górzystym terenie, gdzie wiele obszarów znajduje się poza zasięgiem obserwacji.

FireFinder to innowacyjny systemu do wczesnego wykrywania pożarów w lasach oparty na sieci niedrogich, autonomicznych detektorów. Każdy taki detektor wyposażony jest w kilka sensorów pozwalających na wykrycie gazów lub pyłów emitowanych w trakcie pożaru w promieniu kilkuset metrów. Detektory te zasilane są za pomocą paneli fotowoltaicznych oraz zapasowego akumulatora zapewniającego nieprzerwaną pracę. Każdy detektor wyposażony jest w moduł komunikacyjny sieci LPWAN (NBIoT lub LoRaWAN), która charakteryzuje się niewielkim zużyciem energii, dalekim zasięgiem oraz niskim kosztem modemów. Te aspekty sprawiają, że idealnie nadaje się ona do zastosowania w naszych detektorach. Dane zbierane przez każdy z sensorów są wstępnie analizowane w urządzeniu w oparciu o stworzony algorytm detekcji anomalii, a po wykryciu zaburzenia wysyłane są do centralnego serwera. Tam dane uzyskane z kilku okolicznych detektorów są przetwarzane przez zaawansowany system oparty na metodach sztucznej inteligencji i na tej podstawie zostaje potwierdzony lub odrzucony alarm oraz określana jest estymowana lokalizacja pożaru. Następnie odpowiedni alert jest automatycznie wysyłany do odpowiednich służb. To wszystko przekłada się na szybsze ugaszenie pożaru.

2. SYSTEM PREDYKCJI NASŁONECZNIENIA

Zarządzanie energią jest bardzo ważne dla systemu FireFinder. Nasze sensory są niewielkie, wyposażone w panel fotowoltaiczny o powierzchni 25 cm^2 i instalowane w zacienionym, leśnym środowisku. Częste pomiary są ważne dla szybkiego wykrycia alarmu, jednak odpowiednie gospodarowanie dostępną energią jest kluczowe dla zapewnienia nieprzerwanego funkcjonowania systemu.

W celu lepszego zarządzania częstotliwością transmisji i pomiarów opracowaliśmy system predykcji nasłonecznienia i energii generowanej przez panel PV. Opracowany algorytm predykcji nasłonecznienia określa ile energii urządzenie może zużyć na funkcjonowanie bez ryzyka utraty płynności energetycznej w przyszłości. Nasza koncepcja wykorzystuje informację o poziomie nasłonecznienia. Wykorzystujemy indeks DNI (ang. Direct Normal Irradiance) dla zachmurzonego nieba (standardowy indeks DNI dla nieba bezchmurnego skorygowany o informacje z map radarowych) [3]. Jest to ilość promieniowania słonecznego odbieranego na jednostkę powierzchni przez powierzchnię, która jest zawsze prostopadła do promieni padających w linii prostej ze słońca w jego aktualnym położeniu na niebie. Ta wielkość jest szczególnie interesująca w przypadku skupiających instalacji solarnych i instalacji śledzących położenie słońca. W naszym przypadku zakładamy, że nie znamy dokładnego położenia panelu słonecznego. Dokładne zapisywanie kierunku i ułożenia sensora podczas instalacji zwiększałoby koszty. Dodatkwo te wartości mogą się zmienić, np. przez poluzowanie paska montażowego.

Wartość DNI jest dla nas najbardziej użyteczna, ponieważ reprezentuje maksymalną możliwą do uzyskania wartość nasłonecznienia, przy pewnym ułożeniu, prostopadłym do linii promieni padających w linii prostej od słońca. Następnie wartości DNI korelujemy w czasie z historycznymi wartościami zarejestrowanego uzysku energetycznego na panelu fotowoltaicznym. W ten sposób znajdujemy pory dnia, kiedy panel fotowoltaiczny silnie reaguje na słońce. Zwłaszcza w środowisku leśnym, w którym panuje silne zacienienie zmieniające się w czasie ze zmianami pór roku oraz ze wzrostem drzewostanów, bezpośrednie słońce pada na nasze urządzenia w różnych porach dnia. Kiedy zidentyfikujemy jakie pory dania są najbardziej wartościowe dla procesu ładowania sensora, przy predykcji dostępności energetycznej na najbliższe dni, odczyty DNI dla tych godzin dnia są uwzględniane z większą wagą.

Ponieważ kąty padania słońca zmieniają się w ciągu roku, jak i drzewostan może się zmienić, przy wyliczaniu korelacji czasowej DNI do ładowania paneli wykorzystujemy N (ok. 14) ostatnich dni, z których wybieramy najbardziej słoneczne, eliminując w ten sposób efekt chwilowego zachmurzenia.

Dokładny opis wyliczenia predykcji energii generowanej przez panel fotowoltaiczny określa poniższy algorytm:

1. Wczytanie danych historycznych z ostatnich N (= 14) dni o nasłonecznieniu lokalizacji instalacji sensora. Wykorzystywany jest index DNI dla nieba zachmurzonego.
2. Wczytanie danych historycznych z ostatnich N (= 14) dni, o poziomach naładowania baterii, czasie kiedy bateria była w stanie ładowania i wartości prądu ładowania dla konkretnego sensora, dla którego wyliczana jest predykcja energii generowanej przez panel fotowoltaiczny.
3. Do wyliczeń odrzucane jest $k\%$ (= 0.5) dni z najmniejszą zsumowaną z całego dnia wartością indeksu DNI, odrzucamy dni pochmurne, żeby odfiltrować zachmurzenie od rzeczywistego zacienienia spowodowanego topograficzną lokalizacją urządzeń.
4. Dla każdej godziny wyliczana jest istotność poziomu nasłonecznienia zadana wzorem, gdzie WR_h to waga przyporządkowana do godziny h, która określa potencjał naładowania urządzenia w danym oknie czasowym. D to ostatnie historyczne dni, które nie zostały odfiltrowane, DNI_{dh} to wartość indeksu DNI w dniu d, godzinie h, a EPV_{dh} to wartość energii wygenerowanej przez panel fotowoltaiczny w dniu d, godzinie h.
$$WR_h = \frac{1}{D}\sum_{d}^{D}\frac{EPV_{dh}}{DNI_{dh}} \tag{1}$$
5. Dysponując wyliczonymi wagami WR_h dla wszystkich godzin i znając prognozę indeksu DNI, którą otrzymujemy z tego samego źródła co dane historyczne, czyli z API Open Weather Maps, możemy wyliczyć przewidywaną wartość energii, która zostanie wygenerowana przez panel PV w najbliższych dniach. EPV to przewidywana wartość energii, która zostanie wygenerowana przez panel PV w najbliższych D_f dniach, a DNI_{dh} to prognozowana wartość indeksu DNI w dniu d, godzinie h.
$$EPV = \sum_{d}^{D_f}\sum_{h}^{H_f} WR_h DNI_{dh} \tag{2}$$

Opracowane zostały również algorytmy dobierające optymalny energetycznie cykl pracy urządzenia w zależności od informacji uzyskanych z systemu predykcji nasłonecznienia i energii generowanej przez panel PV oraz obecnego poziomu zagrożenia pożarowego wyznaczanego z użyciem metody opracowanej przez Instytut Badawczy Leśnictwa [1, 2]. Rysunek 2 przedstawia schemat blokowy opracowanego algorytmu do kontrolowania częstotliwości wykonywania pomiarów oraz wysyłania danych na serwer. Ponieważ konieczność wysyłania danych pomiarowych na serwer wymaga dodatkowego użycia energii, pracę urządzeń możemy podzielić na dwa zasadnicze tryby. Pierwszy z nich zakłada wysyłanie danych pomiarowych co N minut, gdzie dokładna wartość częstotliwości ustalana jest na podstawie powyżej opisanych czynników. Taki tryb działania ma uzasadnienie, kiedy chcemy gromadzić dane środowiskowe, których wartość wykracza poza obszar samego wykrywania pożarów, np. w celu prowadzenie badań środowiskowych w lasach. Są jednak sytuacje, kiedy nie chcemy wysyłać niepotrzebnych (niezawierających anomalii) danych, żeby ograniczyć zużycie energii i wysyłać na serwer tylko notyfikację alarmową. W drugim trybie "*only alert send*" urządzenie prowadzi pomiary z zadaną częstotliwością, ale nie wysyła danych na serwer. Po stronie urządzenia zaimplementowaliśmy prosty algorytm wyliczający wartość alarmu i tylko jeżeli wyliczona wartość przekroczy pewną zadaną wartość progową urządzenie wysyła dane, notyfikując serwer o wystąpieniu potencjalnego alarmu. Funkcja alarmu dana jest poniższym równaniem:
$$f_{alarm} = \frac{voc^a * abs(voc_d)}{abs(hum_{d5m})^b}$$
$$SEND\ ALARM\ \ if\ \ f_{alarm} >= T_a \tag{3}$$

gdzie voc to wartość wyliczana przez czujnik stężenia związków lotnych, voc_d to pochodna tej wartości z przyrostem licznika zmian równym 1 (następny pomiar minus poprzedni pomiar), hum_{d5m} to średnia ruchoma uzyskanych wartości wilgotności powietrza, gdzie okno uśredniania wynosi 5 pomiarów. Jeśli obliczona maksymalna wartość funkcji alarmowej przekroczy pewną wartość progową, to wtedy ryzyko wykrycia pożaru jest potwierdzane w urządzeniu wykrywającym oraz generowany i wysyłany jest sygnał alarmowy do serwera. Wartości potęg a i b oraz wartość progowa T_a to parametry sterujące równaniem. Analizowaliśmy różne kombinacje monitorowanych parametrów środowiskowych. Największy problem występował z pojawiającymi się fałszywymi alarmami (duża wartość voc), kiedy dochodziło do skoków wartości wilgotności. Nie bazujemy na samych wartościach bezwzględnych, ponieważ z czasem czujniki mogą ulegać dekalibracji i mieć różną wartość tła. Dlatego znacznie lepiej sprawuje się wykorzystanie analizy pochodnych, która umożliwia monitorowanie szybkości zmian. Wprowadzenie czynnika kompensacyjnego, doprowadziło do opracowania równania, które jest stabilnym predykatywem wystąpienia zagrożenia pożarowego.

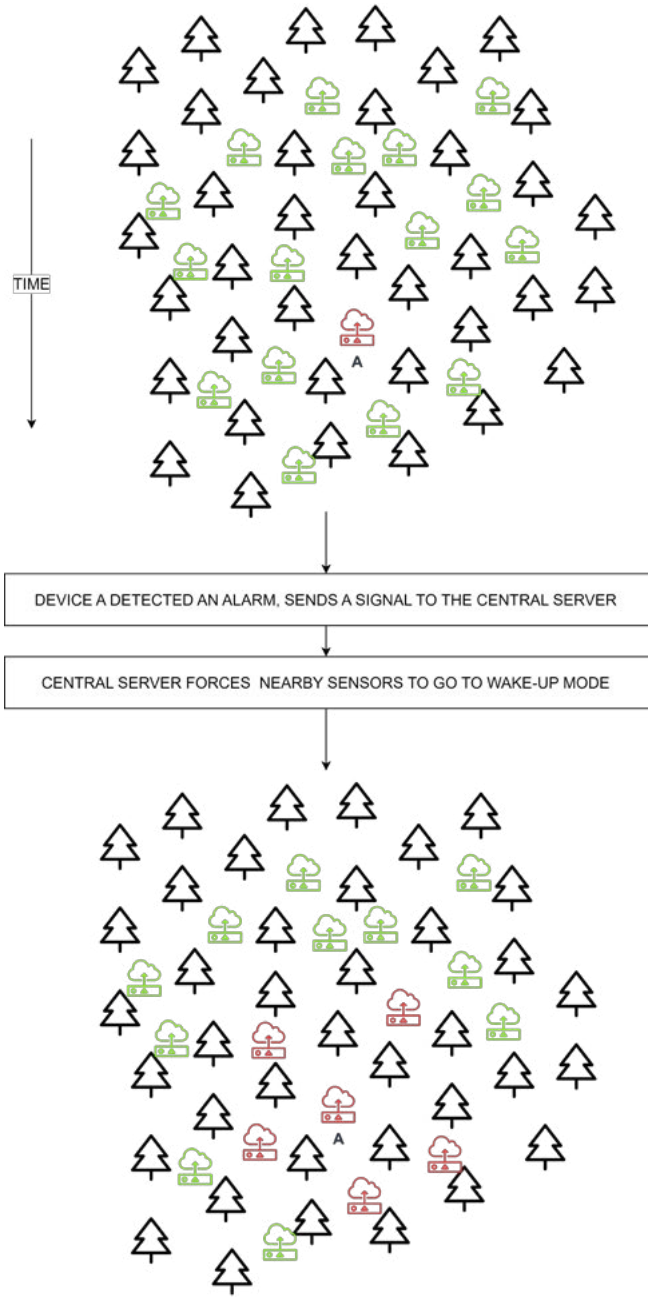
3. PROCEDURA DETEKCJI

Opracowaliśmy algorytm umożliwiający weryfikację alarmu oraz wstępne określenie lokalizacji i tempa rozprzestrzeniania się pożaru. Dane wejściowe stanowią szeregi czasowe reprezentujące wartości zmierzone przez poszczególne urządzenia wraz z przypisanymi im lokalizacjami urządzeń. Dodatkowym wejściem danych jest informacja o aktualnym stanie pogodowym (prędkość i kierunek wiatru, ciśnienie atmosferyczne, zachmurzenie, nasłonecznienie, temperatura i wilgotność powietrza) w celu lepszej analizy stanu atmosfery. Wyjściem algorytmu jest prawdopodobieństwo wykrycia pożaru oraz jego estymowana lokalizacja.

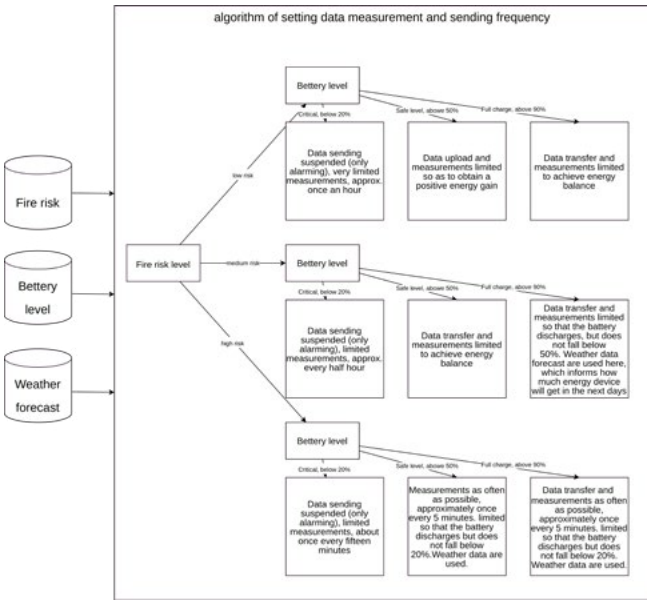
Sam proces detekcji alarmu jest podzielony na dwie logicznie i fizycznie separowalne fazy. Faza pierwsza, prealarm, dokonywana jest na urządzeniu końcowym. Każdy pojedynczy czujnik FireFinder stale monitoruje parametry środowiskowe i oblicza prostą funkcję wartości alarmu, której parametrami są wartości monitorowanych zmian oraz pochodne tych wartości.

W momencie, gdy wartość funkcji alarmu przekroczy pewien próg, czujnik FireFinder wysyła do serwera notyfikację o wykryciu alarmu. W zależności od trybu pracy czujnik może wysyłać monitorowane parametry w sposób ciągły lub być w trybie czuwania i wysyłać dane tylko, kiedy wykryje alarm. Po wykryciu alarmu czujnik przez pewien czas wchodzi w tryb zwiększonej częstotliwości dokonywania pomiarów. Kiedy serwer otrzyma informację o alarmie od pojedynczego czujnika, natychmiast wysyła sygnał wzbudzenia do sąsiadujących czujników FireFinder, celem zwiększenia częstotliwości wykonywania i wysyłania pomiarów. Następnie serwer rozpoczyna fazę drugą detekcji alarmu, tj. przetwarzania danych pochodzących z wielu sensorów zlokalizowanych w okolicy czujnika, który wykrył alarm. Procedura ta ma na celu ewentualne wyeliminowanie fałszywego alarmu tj. określenie prawdopodobieństwa wystąpienia pożaru, oraz określenia lokalizacji pożaru. Rysunek 1 pokazuje ogólny schemat aktywacji sieci czujników FireFinder. W pierwszej części sensor A wykrywa alarm, następnie informuje o tym serwer, w konsekwencji, serwer wzbudza sąsiadujące czujniki do trybu częstego mierzenia i wysyłania danych.

Schemat działania modelu uczenia maszynowego przedstawiono na rysunku 3. Wykorzystanie współdzielonych wag razem z użyciem warstwy sumującej, umożliwiło nam trenowanie i wnioskowanie modelu, który analizuje dane z nieokreślonej sztywno liczby urządzeń.



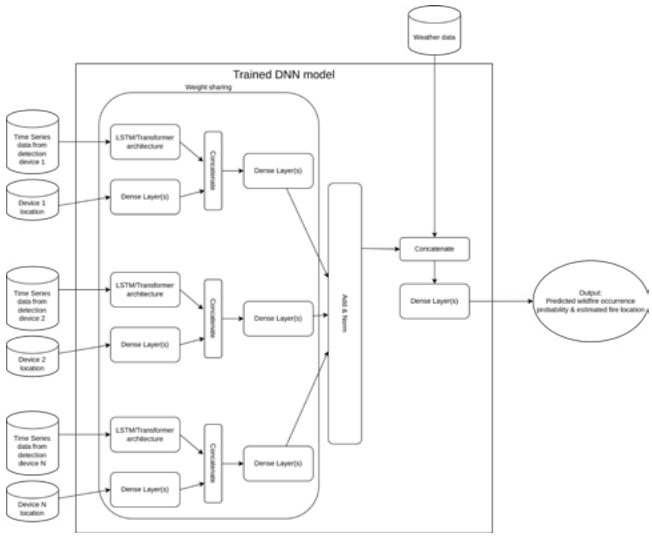
Rysunek 1: Schemat aktywacji sieci czujników FireFinder. W pierwszym etapie **(górny rysunek)** pojedynczy sensor A wykrywa wstępny alarm, następnie wysyła notyfikację o wystąpieniu alarmu na serwer. Serwer wzbudza sąsiadujące czujniki do trybu częstego mierzenia i wysyłania danych **(dolny rysunek)**



Rysunek 2: Algorytm sterowania częstotliwością wykonywania pomiarów i wysyłania danych na serwer. Algorytm wykorzystuje trzy informacje jako dane wsadowe: poziom zagrożenia pożarowego, obecny poziom naładowania baterii oraz prognozę pogody (dokładniej opisany wcześniej współczynnik DNI i oszacowaną na jego podstawie przewidywaną wartość energii, która zostanie wygenerowana przez panel PV w najbliższych dniach). Częstotliwość pomiarów jest wzmacniana przez wysokie zagrożenie pożarowe, wysoki poziom naładowania baterii i korzystną prognozę przyszłego ładowania baterii. Czynnikami hamującymi częstotliwość wykonywania pomiarów są: niskie zagrożenie pożarowe, niski poziom naładowania baterii oraz niska wartość przewidywanej wartość energii, która zostanie wygenerowana przez panel PV w najbliższych dniach.

4. PODSUMOWANIE

W niniejszym opisie przedstawiliśmy mechanizm działania systemu FireFinder, który wykorzystuje technologię internetu rzeczy do zdalnego monitorowania lasów w celu umożliwienia szybkiej detekcji dzikich pożarów i tym samym natychmiastowej reakcji służb gaśniczych. Czujniki komunikują się z centralnym serwerem za pomocą sieci Narrowband IoT lub LoRaWAN. Parametry środowiskowe takie jak temperatura, wilgotność czy koncentracja związków lotnych powstałych na skutek spalania materii organicznej są stale monitorowane. Kiedy urządzenie wykryje wstępny alarm, wysyła notyfikacje na serwer. Oprogramowanie w chmurze wzbudza sąsiednie sensory do częstszego wysyłania danych pomiarowych i na podstawie danych pochodzących z kilku detektorów wykluczane są fałszywe alarmy oraz estymowana jest lokalizacja pożaru. Te informacje przesyłane są niezwłocznie w formie alarmu do odpowiednich służb. Tak skonstruowany scenariusz działania umożliwia znaczne skrócenie czasu od zaprószenia ognia do przybycia służb gaśniczych, znacznie redukując tym samym szkody. W ramach dalszych prac, planujemy przeprowadzać pełnoskalowe wdrożenia systemu, dzięki który zgromadzimy rzeczywiste dane umożliwiające dotrenowanie i pełniejszą walidację modeli jak i całego systemu.



Rysunek 3: Architektura zastosowanego modelu uczenia maszynowego, do określania prawdopodobieństwa wykrycia pożaru oraz estymacji lokalizacji pożaru. Do przetwarzania danych z sensorów stosowano warstwy LSTM (Long short-term memory) [5] lub sieci o architekturze transformer [4]. Na syntetycznym zbiorze testowym zawierającym 10 tys. przykładów (po 50% dla przykładów zawierających "pożar" i "brak zagrożenia"). Wartość dokładności klasyfikacji wyniosła 94.5%, uzyskanie danych ewaluacyjnych pochodzących ze środowiska rzeczywistego nastąpi po przeprowadzenie pełnoskalowych wdrożeń.

LITERATURA

[1] Opis metody obliczania zagrożenia pożarowego opracowany przez Instytut Badawczy Leśnictwa: https://bazapozarow.ibles.pl/zagrozenie/metoda_IBL_nowa.pdf

[2] Rozporządzenie Ministra Środowiska w sprawie szczegółowych zasad zabezpieczenia przeciwpożarowego lasów: https://isap.sejm.gov.pl/isap.nsf/download.xsp/WDU20060580405/O/D20060405.pdf

[3] Strona internetowa Open Weather Map zawierająca opis wyliczania indeksu DNI, https://openweathermap.org/api/solar-radiation/behind-solar-radiation-api

[4] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[5] Karim, Fazle, et al. "Multivariate LSTM-FCNs for time series classification." Neural networks 116 (2019): 237-245.

ANALIZA SKUTECZNOŚCI DETEKCJI SYGNAŁÓW GMSK W KANALE Z PROPAGACJĄ WIELODROGOWĄ PRZY UŻYCIU METOD GŁĘBOKIEGO UCZENIA

EFFICIENCY ANALYSIS OF GMSK SIGNAL DETECTION IN MULTIPATH CHANNEL BASED ON DEEP LEARNING METHODS

Alicja Olejniczak[1]; Krzysztof K. Cwalina[2]; Jarosław Sadowski[3]; Jacek Stefański[4]

[1] Politechnika Gdańska, Wydział Elektroniki Telekomunikacji i Informatyki, Gdańsk, alicja.olejniczak@pg.edu.pl
[2] Politechnika Gdańska, Wydział Elektroniki Telekomunikacji i Informatyki, Gdańsk, kkcwalina@eti.pg.edu.pl
[3] Politechnika Gdańska, Wydział Elektroniki Telekomunikacji i Informatyki, Gdańsk, jaroslaw.sadowski@eti.pg.edu.pl
[4] Politechnika Gdańska, Wydział Elektroniki Telekomunikacji i Informatyki, Gdańsk, jacek.stefanski1@pg.edu.pl

DOI: 10.15199/59.2023.4.50

Streszczenie: **Odbiór sygnałów radiowych w środowisku wewnątrzbudynkowym jest istotnym problemem współczesnej radiokomunikacji. W celu zwiększenia skuteczności istniejących metod odbioru radiowego coraz częściej wykorzystuje się narzędzia z obszaru uczenia maszynowego. Niniejszy artykuł prezentuje analizę skuteczności głębokiej sieci neuronowej w odbiorze sygnałów GMSK w kanale z zanikami i propagacją wielodrogową. Prezentowane wyniki porównane zostały z klasycznym detektorem MLSE, a miarą skuteczności była bitowa stopa błędów.**

Abstract: **Receiving radio signals in an indoor environment is an important aspect of modern radio communication. In order to increase the effectiveness and reliability of the existing reception solutions, tools from the field of machine learning are increasingly used. This article presents an analysis of the effectiveness of a deep neural network in reception of GMSK signals in the multipath, fading channel. The presented results were compared with the classical MLSE detector and the measure of effectiveness was the bit error rate.**

Słowa kluczowe: **EPA, głębokie uczenie, GMSK, sztuczna inteligencja.**

Keywords: **artificial intelligence, deep learning, EPA, GMSK.**

1. WSTĘP

Dynamiczny rozwój współczesnych systemów radiokomunikacyjnych determinuje potrzebę wprowadzania coraz to nowych rozwiązań przystających zarówno do wymagań standaryzacyjnych i technologicznych, jak i tych związanych z szeroko rozumianym QoS (ang. *Quality of Service*). Nowe technologie, jak np. IoT (ang. *Internet of Things*) czy 5G niosą za sobą szereg wyzwań i problemów dotyczących transmisji bezprzewodowej w ogólności – od formowania sygnałów radiowych w warstwie fizycznej, poprzez integrację protokolarną i alokację zasobów, po realizację produktów końcowych o niskich kosztach obliczeniowych i produkcyjnych. Zapewnienie wysokiej jakości pracy w czasie rzeczywistym, szczególnie w warunkach trudnych, jak np. środowisko wewnątrzbudynkowe, w którym zjawiska takie jak propagacja wielodrogowa czy zaniki negatywnie wpływają na transmisję sygnałów radiowych, wymaga stosowania wydajnych obliczeniowo rozwiązań, które umożliwiają analizę problemów o charakterze nieliniowym. Jednym z obszarów, który pozwala dostarczyć narzędzia o wspomnianych możliwościach jest niewątpliwie sztuczna inteligencja, a w szczególności głębokie uczenie.

W dostępnej literaturze znaleźć można wiele przykładów wykorzystania modeli uczenia maszynowego do odbioru sygnałów radiowych. W [12] zaprezentowano detekcję symboli OFDM (ang. *Orthogonal Frequency-Division Multiplexing*) w kanale WINNER II poprzez sieć DNN (ang. *Deep Neural Network*). Skuteczność modelu porównano z metodą MMSE (ang. *Minimum Mean-Square Error*). Dodatkowo, przeanalizowano wpływ liczby przesyłanych sygnałów pilotowych na końcową skuteczność modelu.

W [8] autorzy przedstawili strukturę 1-bitowego odbiornika OFDM bazującego na dwóch sieciach neuronowych realizujących kolejno estymację kanału oraz detekcję transmitowanych symboli. Prezentowane wyniki porównane zostały z metodami estymacji LS (ang. *Least-Square*) i ZF (ang. *Zero-Forcing*).

Ewaluację głębokiej sieci neuronowej przedstawiono również w [6]. Prezentowane badania objęły transmisję w paśmie terahercowym. Problem detekcji przedstawiony został jako problem klasyfikacyjny, a końcowe wyniki bitowej stopy błędów porównano z algorytmem MMSE (ang. *Minimum Mean-Square Error*). Definicję zagadnienia odbioru pakietów w postaci problemu klasyfikacyjnego odnaleźć można w wielu badaniach, m.in. [3, 10, 13]. Należy jednak mieć na uwadze, że podejście to może się okazać kłopotliwe w przypadku analizy pakietów o dużej długości.

Liczne prace naukowe rozszerzają analizę detekcji sygnałów radiowych do systemu tzw. end-to-end, w którym zestaw kilku modeli uczenia maszynowego realizuje funkcje toru nadawczego i odbiorczego, jak również samego kanału radiowego. W [14] zaprezentowano system realizujący zarówno modulację jak i detekcję sygnałów radiowych dla kanału z zanikami. Prezentowane wyniki porównano z bitową stopą błędów dla metody MMSE i modulacji 64-QAM (ang. Quadrature Amplitude Modulation). W [4] przeanalizowano system end-to-end, w którym trening modeli zrealizowano w dwóch fazach: pierwsza bazowała na ogólnym modelu kanału, zaś w drugiej następowało douczenie modeli w oparciu o rzeczywisty

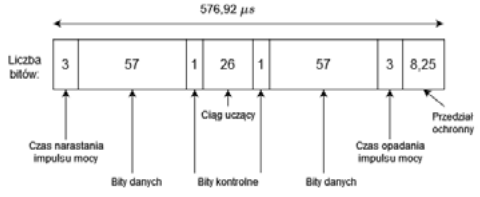
kanał transmisyjny. Warto zwrócić uwagę, iż objęcie uczeniem maszynowym części nadawczej i odbiorczej może prowadzić do uproszczenia operacji realizowanych w ramach warstwy fizycznej, jednak równocześnie uniemożliwia zastosowanie takiego rozwiązania w strukturze protokolarnej istniejących systemów radiokomunikacyjnych.

Pomimo dużej liczby prac analizujących zagadnienia z zakresu odbioru radiowego, większość z nich obejmuje technikę OFDM oraz modulacje fazy i częstotliwości. Pojedynczy artykuł prezentuje analizę skuteczności modeli uczenia maszynowego w detekcji sygnałów z modulacją GMSK (ang. Gaussian Minimum Shift Keying) [15], dotyczy jednak transmisji w kosmosie.

Mając na uwadze ograniczenia wyróżnione w przeprowadzonym przeglądzie literaturowym, w niniejszym artykule podjęto zagadnienie analizy skuteczności głębokiego uczenia w detekcji sygnałów GMSK dla kanału z zanikami. Prezentowany model głębokiej sieci neuronowej realizuje jedynie funkcje części odbiorczej, nie ingerując tym samym w protokolarną strukturę transmitowanych pakietów. W rozdziale drugim przedstawiony zostanie opis części symulacyjnej realizującej generację sygnałów GMSK oraz kanału z zanikami. W rozdziale trzecim zaprezentowany zostanie proces projektowania i ewaluacji modelu głębokiego uczenia wraz z uzyskanymi wynikami detekcji GMSK. Ostatni rozdział stanowić będzie podsumowanie omówionych prac badawczych.

2. ŚRODOWISKO SYMULACYJNE

W pierwszym etapie realizowanych prac badawczych przygotowane zostały zestawy danych treningowych i testowych w celu ewaluacji modeli głębokiego uczenia. Bloki danych informacyjnych wraz ciągami uczącymi uporządkowane zostały w formie pakietów TCH/F (ang. *Traffic Channel/Fullrate*) według specyfikacji systemu GSM (ang. Global System for Mobile Communications) [11], których strukturę zaprezentowano na Rys. 1 (uzasadnienie wyboru niniejszej struktury można odnaleźć w [9]).



Rys. 1. Struktura pakietu w kanale TCH/F [11]

Następnie, sekwencje bitów przetwarzane były w ramach modulatora GMSK, zgodnie ze wzorem [5]:

$$s(t) = \sum_{k=0}^{K-1} s_k(t), \qquad (1)$$

gdzie $s(t)$ reprezentuje K zmodulowanych segmentów, a k-ty segment o czasie trwania T_b można opisać jako:

$$s_k(t) = \begin{cases} A_0 \sin \Phi_k(t) \text{ dla } (k-\frac{1}{2})T_b \le t \le (k+\frac{1}{2})T_b \\ 0 \quad \text{dla pozostałych } t \end{cases}. \qquad (2)$$

Amplituda i opisana funkcją gaussowską faza chwilowa sygnału to kolejno: A_0 oraz $\Phi_k(t)$.

Ostatnim elementem symulatora generującego dane uczące było modelowanie kanału radiowego. W ramach prezentowanych badań przyjęto kanał z propagacją wielodrogową EPA (ang. *Extended Pedestrian A*) zdefiniowany w standardzie 3GPP [1].

Tab. 1. Parametry modelu EPA [1]

Względne opóźnienie składowej sygnału [ns]	*Względna moc składowej sygnału [dB]*
0	0
30	-1
70	-2
90	-3
110	-8
190	-17,2
410	-20,8

Pełen zbiór danych treningowych składał się z 40 bloków dla 5 różnych ciągów uczących, a każdy blok zawierał 2004 pakiety. Dodatkowo, do wszystkich danych dodano szum AWGN w zakresie E_b/N_0 = 1-18 dB. Tym samym, dla kanału EPA wygenerowanych zostało ponad 7 000 000 pakietów. Niestacjonarność kanału uwzględniona w wykorzystanym zbiorze treningowym umożliwiła ewaluację modelu w warunkach zbliżonych do warunków rzeczywistych. Należy również zauważyć, iż proponowane rozwiązanie może odnaleźć zastosowanie przede wszystkim w systemach IoT w warunkach trudnych (np. deep indoor), gdzie nieruchome modemy IoT będą determinować relatywnie długie czasy stacjonarności.

3. DETEKCJA SYGNAŁÓW GMSK

Główną ideą działania modeli uczenia maszynowego jest reprezentacja złożonych (również nieliniowych) zależności w postaci prostych operacji matematycznych [5].

Jako strukturę DL (ang. *Deep Learning*) pełniącą funkcję detektora GMSK wybrano sieć splotową CNN (ang. *Convolutional Neural Network*), której skuteczność potwierdziły badania prezentowane w [9]. Proces uczenia sieci neuronowej sprowadza się do modyfikacji parametrów sieci, tak aby dla danych wejściowych ze zbioru uczącego, dane na wyjściu sieci były jak najbardziej zbliżone do wektora referencyjnego. W niniejszych badaniach za metrykę porównawczą przyjęto funkcję straty BCE (ang. *Binary Cross Entrophy*), którą można opisać jako:

$$BCE = -\frac{1}{m}\sum_{i=1}^{m}(y_i \log y_i + (1-y_i)\cdot \log(1-y_i)), \qquad (3)$$

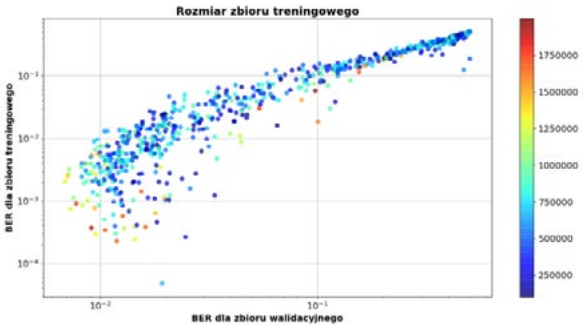
gdzie y_i oraz $\hat{y}_i$ należy traktować kolejno jako i-tą wartość referencyjną i predykcję sieci. Mając na uwadze, że wartości funkcji straty wyrażone są w postaci prawdopodobieństwa wystąpienia błędu binarnego, na wyjściu sieci we wszystkich testowanych modelach zastosowano warstwę sigmoidalną, skalującą dane wyjściowe do zakresu [0,1].

W kolejnym etapie ewaluacji, poszukiwano najefektywniejszej architektury dla detektora DL CNN. Z uwagi na szeroki zakres parametrów architektury sieci i parametrów związanych z procesem uczenia (m.in. kernel, liczba iteracji) zastosowano metodę poszukiwania losowego (ang. *random search*) [7]. Zakres testowanych hiperparametrów (z wyjątkiem liczby warstw i funkcji aktywacji) przedstawiono w Tab.2.

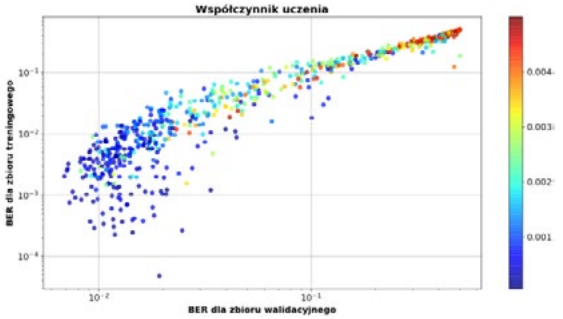
parametr	*zakres*	*rozkład*
batch_size	[1,1000]	int_uniform
wsp. uczenia	[0,00009; 0,005]	uniform
rozmiar zbioru danych	[100 000; 2 000 000]	int_uniform
kernel	[2; 9]	int_uniform
l. iteracji	{400}	-
dropout	[0,1; 0,5]	uniform

Parametr batch_size należy traktować jako zbiór danych treningowych, po przetworzeniu których następuje aktualizacja parametrów sieci, a współczynnik uczenia jako krok tejże aktualizacji. Kernel determinuje rozmiar okna, dla którego realizowana jest operacja splotu wewnątrz sieci CNN. Czas trwania fazy uczenia definiuje liczba iteracji (ang. *epochs*). Dodatkowo, w celu poprawy generalizacji sieci, zastosowano warstwę dropout, realizującą losową dezaktywację neuronów podczas procesu treningu.

Nie wszystkie parametry wykazywały jednoznaczny wpływ na końcową skuteczność detekcji GMSK, co jest zjawiskiem w pełni uzasadnionym. Efektywność modelu determinuje bowiem nie tylko zestaw hiperparametrów jako całość, ale również sam przebieg procesu uczenia, który nie jest stały nawet w przypadku ewaluacji tej samej struktury kilkukrotnie (zbiór danych treningowych i parametrów wejściowych sieci jest dobierany losowo).



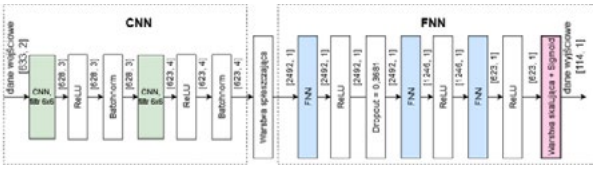
Rys. 2. Wpływ rozmiaru zbioru treningowego na końcową skuteczność detekcji



Rys. 3. Wpływ współczynnika uczenia na końcową skuteczność detekcji

Niemniej jednak, możliwe było wyodrębnienie korelacji pomiędzy końcową skutecznością detekcji GMSK (zarówno dla zbioru treningowego, jak i walidacyjnego) a rozmiarem zbioru uczącego i wielkością współczynnika uczenia, co zaprezentowano na rysunkach 2 i 3. Należy zauważyć, że najwyższą skuteczność detekcji uzyskano dla modeli z niskim współczynnikiem uczenia, dla dużych zbiorów danych treningowych.

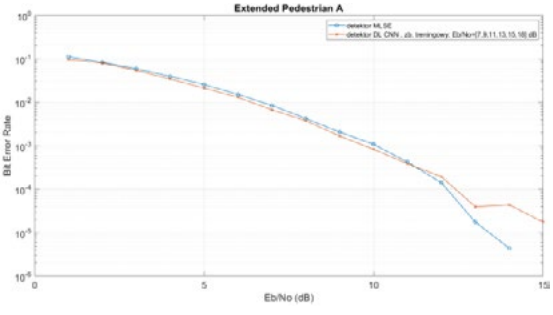
Architekturę najlepszego z testowanych modeli zaprezentowano na Rys. 4. Strukturę podzielono na dwie części: 2-warstwową sieć splotową wraz z warstwami normalizującymi oraz klasyczną sieć neuronową, na którą składały się 3 warstwy i warstwa skalująca dane do zakresu [0,1]. Dla wszystkich warstw zastosowano funkcję aktywacji ReLU (ang. Rectified Linear Unit).



Rys. 4. Najlepsza struktura detektora DL CNN

W tym miejscu należy wspomnieć, iż ewaluacja modeli neuronowych nie odbywała się dla pełnego zakresu E_b/N_0. Zgodnie z wnioskami prezentowanymi w [9] oraz [2], zbytnie zniekształcenie próbek sygnału GMSK w zbiorze treningowym, wpływa negatywnie na skuteczność detekcji i generalizację modelu. Dlatego też podczas fazy treningu ujęto dane dla $E_b/N_0 > 6$ dB.

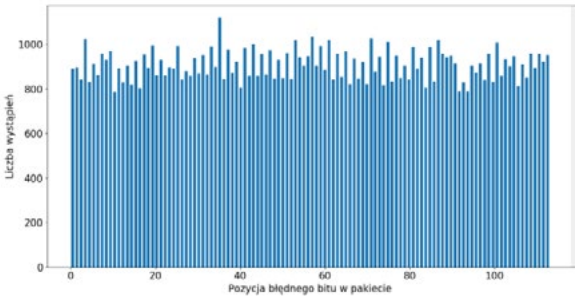
Charakterystyki BER detektora DL CNN oraz optymalnego detektora MLSE dla zbioru danych testowych przedstawiono na Rys. 5.



Rys. 5. Charakterystyki BER detekcji sygnałów GMSK

Można zauważyć, iż detektor DL CNN osiągnął skuteczność wyższą od detektora optymalnego w zakresie E_b/N_0 =1-12 dB, niższą w zakresie 12-15 dB i taką samą w zakresie 15-18 dB. Szczególne pogorszenie skuteczności zauważyć można dla poziomu 14 dB, czyli wartości nieuwzględnionej w zbiorze danych treningowych. Niestety, pomimo rozszerzenia zakresu E_b/N_0 w zbiorze uczącym, nie udało się uzyskać skuteczności wyższej, niż ta prezentowana na Rys. 5. Pomimo pogorszenia efektywności sieci dla niższego poziomu szumu, w zakresie 1-7 dB, który nie został ujęty w zbiorze treningowym, detektor DL CNN wciąż był w stanie odtwarzać dane ze skutecznością zbliżoną do detektora MLSE.

Analiza rozkładu błędów występujących w pakiecie dla przykładowych bloków informacyjnych pozwoliła wykluczyć występowanie błędów systematycznych, co przedstawiono na Rys. 6.



Rys. 6. Liczba wystąpień błędów na kolejnych pozycjach pakietu TCH/F

4. PODSUMOWANIE

Prezentowane wyniki prac, będących kontynuacją badań przedstawionych w [9], pozwoliły potwierdzić możliwość zastosowania modeli głębokiego uczenia do detekcji sygnałów GMSK przy zachowaniu skuteczności zbliżonej do optymalnego detektora MLSE, nawet w przypadku kanału, w którym występuje zjawisko propagacji wielodrogowej. Otwartą kwestią pozostaje wciąż poprawa efektywności modelu DL dla wysokich wartości E_b/N_0. Właściwym wydaje się przypuszczenie, iż wprowadzenie modelu sieci neuronowej z pamięcią, pozwoli na lepsze wyodrębnienie charakteru zakłóceń i zniekształceń sygnału na podstawie próbek sekwencji uczących w pakiecie dostarczonych na wejście sieci. Jest to jednak zagadnienie, które przeanalizowane zostanie w ramach prac prowadzonych w przyszłości.

LITERATURA

[1] 3GPP, TS 136 116. 2015. "LTE Evolved Universal Terrestrial Radio Access (E- UTRA) Relay Radio Transmission and Reception."

[2] Al-Baidhani, Amer, H. Howard Fan. 2019. "Learning for Detection: A Deep Learning Wireless Communication Receiver Over Rayleigh Fading Channels." *2019 International Conference on Computing, Networking and Communications, ICNC 2019*., 6–10.

[3] Dakic, Kosta, Bassel Al Homssi, Akram Al-Hourani, Margaret Lech. 2021. "LoRa Signal Demodulation Using Deep Learning, a Time-Domain Approach." *IEEE Vehicular Technology Conference*. 2021-April: 1–6.

[4] Dörner, Sebastian, Sebastian Cammerer, Jakob Hoydis, Stephan ten Brink. 2017. "Deep Learning-Based Communication over the Air." *ArXiv*. 12 (1): 132–43.

[5] Goodfellow, Ian, Yoshua Bengio, Aaron Courville. 2016. *Deep Learning*. London: MIT Press.

[6] He, Dongxuan, and Zhaocheng Wang. 2022. "Deep Learning-Assisted Demodulation for TeraHertz Communications Under Hybrid Distortions." *IEEE Communications Letters*. 26 (2): 325–29.

[7] Li, Lisha, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, Ameet Talwalkar. 2018. "Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization." *Journal of Machine Learning Research*. 18: 1–52.

[8] Lu, Zan, Li Wei, Youyun Xu. 2019. "An Improved One-Bit OFDM Receiver Based on Model-Driven Deep Learning." *2019 IEEE 5th International Conference on Computer and Communications, ICCC 2019*., 2108–12.

[9] Olejniczak, Alicja, Krzysztof K. Cwalina, Jarosław Sadowski, Jacek Stefański. 2022. "Analiza Efektywności Metod Głębokiego Uczenia w Odbiorze Sygnałów GMSK." *Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne*. 1 (4): 351–54.

[10] Qiao, Yuanjian, Jun Li, Bo He, Wenxin Li, Tongliang Xin. 2020. "A Novel Signal Detection Scheme Based on Adaptive Ensemble Deep Learning Algorithm in SC-FDE Systems." *IEEE Access*. 8: 123514–23.

[11] Stefanski, Jacek. 2000. "Efektywność Odbioru Adaptacyjnego w Systemie Łączności Komórkowej GSM." Politechnika Gdanska.

[12] Ye, Hao, Geoffrey Ye Li, Biing Hwang Fred Juang. 2017. "Power of Deep Learning for Channel Estimation and Signal Detection in OFDM Systems." *IEEE Wireless Communications Letters*. 7 (1): 114–17.

[13] Ye, Hao, Geoffrey Ye Li, Biing Hwang Fred Juang, Kathiravetpillai Sivanesan. 2018. "Channel Agnostic End-to-End Learning Based Communication Systems with Conditional GAN." In *IEEE Globecom Workshops (GC Wkshps)*. Abu Dhabi: IEEE.

[14] Zhu, Banghua, Jintao Wang, Longzhuang He, Jian Song. 2018. "Joint Transceiver Optimization for Wireless Communication PHY with Convolutional Neural Network." *ArXiv*. 37 (6): 1364–73.

[15] Zhu, Manquin, Shengbo Hu, Kai Yang, Tingting Yan, Lv Ye, Hongjuan Li, Yang Jin. 2023. "GMSK Demodulation Combining 1D-CNN and Bi-LSTM Network Over Strong Solar Wind Turbulence Channel." *Radio Science*. 58 (1).

SYSTEM RADIOLOKALIZACYJNY OPARTY NA FARMIE TURBIN WIATROWYCH

RADIOLOCALIZATION SYSTEM BASED ON WIND TURBINE FARM

Jacek Stefański[1]; Jarosław Sadowski[2]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, jstef@eti.pg.edu.pl
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, jarsad@eti.pg.edu.pl

DOI: 10.15199/59.2023.4.51

Streszczenie: **W referacie zaproponowano wykorzystanie farmy turbin wiatrowych do budowy systemu radiolokalizacyjnego dla obiektów latających (samolotów). Zaproponowane rozwiązanie umożliwia oszacowanie położenia obiektu w przestrzeni, zwłaszcza w ruchu, za pomocą sygnałów lokalizacyjnych nadawanych asynchronicznie z wież turbin wiatrowych. Przedstawiono matematyczną analizę metody oraz wyniki badań symulacyjnych efektywności pracy rozwiązania dla wybranych parametrów systemowych.**

Abstract: **The paper proposes using a wind turbine farm to build a radiolocalization system for flying objects (aircraft). The proposed solution makes it possible to estimate the location of an object in space, especially in motion, using localization signals transmitted asynchronously from wind turbine towers. A mathematical analysis of the method and results of simulation studies of the effectiveness of the solution's operation for selected system parameters are presented.**

Słowa kluczowe: **asynchroniczny system radiolokalizacyjny, metoda ATDOA, farma turbin wiatrowych.**

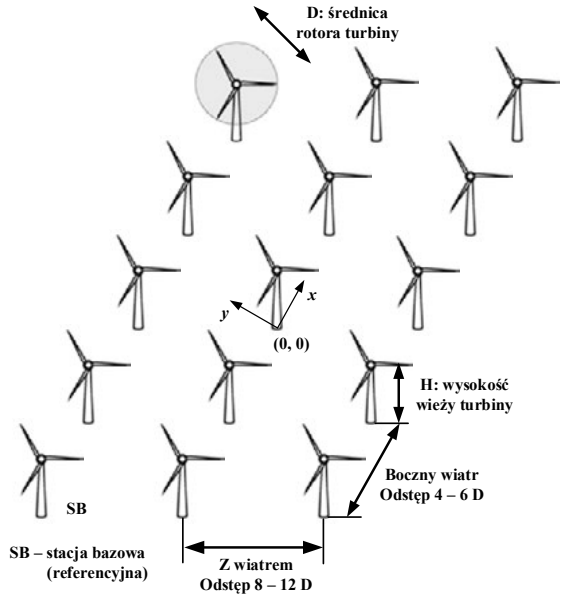
Keywords: **asynchronous radiolocalization system, ATDOA method, wind turbine farm.**

1. WSTĘP

Odnawialne źródła energii rozwijają się na całym świecie znacznie szybciej niż konwencjonalne, a największe znaczenie wśród nich odgrywa energetyka wiatrowa, która opiera się na turbinach wiatrowych. Powszechnie znane są turbiny wiatrowe o poziomej osi obrotu [1]. Z tej grupy najbardziej rozpowszechnioną turbiną jest wirnik trzyłopatowy, stosowany w komercyjnych elektrowniach wiatrowych – farmach wiatrowych. Średnice wirników tego typu turbin wahają się od kilkunastu metrów do ponad 100 metrów, w zależności od mocy nominalnej. W niniejszym artykule autorzy przedstawiają koncepcję budowy systemu radiolokalizacyjnego dla obiektów latających z nadajnikami referencyjnymi umieszczonymi na turbinach farmy wiatrowej. W przedstawionej koncepcji takiego systemu turbiny wiatrowe stanowią miejsce instalacji radiolatarni emitujących radiowe sygnały lokalizacyjne, na podstawie których obiekty latające mogą estymować swoje położenie. Ponadto, wszystkie nadajniki umieszczone na turbinach wiatrowych pracują asynchronicznie, a ich współrzędne nie zmieniają się w czasie eksploatacji systemu. Oznacza to, że każda ze stacji używa lokalnego generatora do nadawania informacji lokalizacyjnych bez potrzeby stosowania synchronizacji elementów sieci [2].

2. STRUKTURA SYSTEMU

Rozważmy przypadek w przestrzeni trójwymiarowej, w której na wieżach turbin wiatrowych z rys. 1 posadowione są nadajniki emitujące asynchronicznie sygnały lokalizacyjne. Przyjęta geometria rozmieszczenia turbin odpowiada zasadom instalacji turbin w rzeczywistych farmach wiatrowych, a odstępy pomiędzy sąsiednimi turbinami zależą od średnicy rotora.



Rys. 1. Przykład farmy wiatrowej dla potrzeb rozważanego systemu radiolokalizacyjnego [3]

Nadajniki, pełniące funkcje stacji referencyjnych (SB) znajdują się w znanych położeniach względem przyjętego układu współrzędnych. Emitowane sygnały lokalizacyjne są odbierane przez obiekt latający (stację ruchomą), np. samolot (SR). Stosując unikatowe identyfikatory w każdym nadajniku, SR może odbierać te sygnały i określać tzw. wirtualne różnice odległości pomiędzy poszczególnymi antenami nadajników i estymować swoje położenie. We wcześniejszych pracach autorów rozważane były przypadki w łączu *w dół*, czyli w kierunku od SR do stacji referencyjnych [4]. Nowością w niniejszym referacie jest rozważanie przypadku w łączu przeciwnym, tzn. od stacji referencyjnych, które skupione są na stosunkowo niewielkim obszarze,

w kierunku obiektu latającego. Warto w tym miejscu podkreślić, że jest to przypadek bardziej złożony w stosunku do wcześniej opisywanego z uwagi na zmianę położenia SR podczas odbioru serii sygnałów referencyjnych – odbiornik lokalizacyjny musi odebrać przynajmniej dwa sygnały z każdej SB w różnych momentach czasu i w różnych położeniach (wymóg ciągłego ruchu stacji ruchomej).

Stacja ruchoma mierzy tzw. wirtualne różnice odległości $D_{i,k}$ dla sygnałów pochodzących od N stacji bazowych

$$D_{i,k} = d_{i,k} + v \cdot \Delta t_k = d_{i,k} + \Delta d_k, \qquad i = 1, \ldots, N \qquad (1)$$

gdzie $d_{i,k}$ reprezentują różnice odległości pomiędzy SR a każdą stacją referencyjną w momentach obserwacji k i k+1 i są opisane następującą zależnością

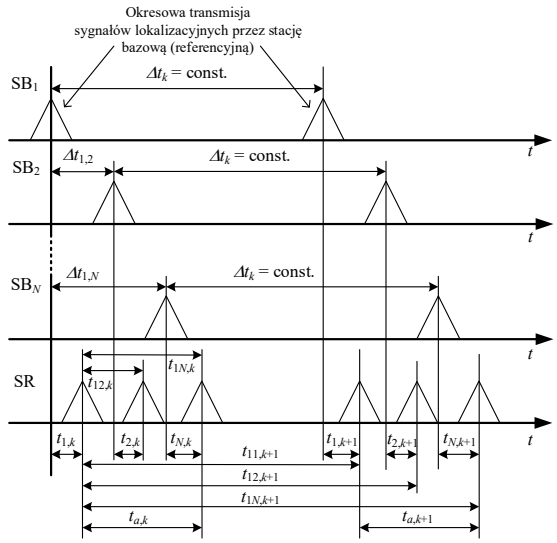
$$\begin{aligned} d_{i,k} &= c \cdot [t_{i,k+1} - t_{i,k}] = r_{i,k+1} - r_{i,k} = \\ &= \sqrt{[X_i - x_{i,k+1}]^2 + [Y_i - y_{i,k+1}]^2 + [Z_i - z_{i,k+1}]^2} + \\ &\quad - \sqrt{[X_i - x_{i,k}]^2 + [Y_i - y_{i,k}]^2 + [Z_i - z_{i,k}]^2} \end{aligned}, \qquad (2)$$

gdzie $t_{i,k}$ i $t_{i,k+1}$ oznaczają czasy propagacji sygnałów radiowych od SB_i do obiektu ruchomego SR odpowiednio w momentach obserwacji k i k+1, (X_i,Y_i,Z_i) reprezentują współrzędne stacji bazowych, natomiast $(x_{i,k},y_{i,k},z_{i,k})$ i $(x_{i,k+1},y_{i,k+1},z_{i,k+1})$ są współrzędnymi SR odpowiednio w momentach obserwacji k i k+1 dla sygnałów pochodzących od i-tej stacji bazowej oraz c jest prędkością rozchodzenia się fal radiowych. Jak wynika z zależności (2), asynchroniczna praca stacji bazowych powoduje odbiór sygnałów radiowych w różnych położeniach SR co znacząco komplikuje proces estymacji jej położenia. W celu rozwiązania układu równań (1) przyjęto, że w czasie pomiarów w dwóch momentach obserwacji k i k+1 stacja ruchoma porusza się po linii prostej w przestrzeni trójwymiarowej. Przy tym założeniu, współrzędne SR w momentach pomiarów wirtualnych różnic odległości można opisać zależnościami:

$$\begin{cases} x_{i,k} = x_{1,k} + t_{1i,k} \cdot v \cdot V_x \\ y_{i,k} = y_{1,k} + t_{1i,k} \cdot v \cdot V_y \\ z_{i,k} = z_{1,k} + t_{1i,k} \cdot v \cdot V_z \end{cases}, \; i = 2, \ldots N, \qquad (3a)$$

$$\begin{cases} x_{i,k+1} = x_{1,k} + t_{1i,k+1} \cdot v \cdot V_x \\ y_{i,k+1} = y_{1,k} + t_{1i,k+1} \cdot v \cdot V_y \\ z_{i,k+1} = z_{1,k} + t_{1i,k+1} \cdot v \cdot V_z \end{cases}, \; i = 1, \ldots N, \qquad (3b)$$

gdzie $(x_{1,k},y_{1,k},z_{1,k})$ reprezentują współrzędne SR w momencie odbioru sygnałów radiowych od SB_1 (zakłada się również, że SB_1 jest stacją odniesienia), V_x, V_y i V_z oznaczają wektory orientacji trajektorii ruchu w przestrzeni trójwymiarowej względem poszczególnych osi układu współrzędnego, natomiast v jest skalarną prędkością poruszania się SR. Rozwiązaniem układu równań (1) są współrzędne stacji ruchomej $(x_{1,k},y_{1,k},z_{1,k})$, wektory orientacji trajektorii V_x, V_y i V_z oraz v. Wynika z tego, że w systemie powinno pracować minimum siedem stacji bazowych. Na rys. 2 przedstawiono ilustrację rozważanego zagadnienia w oparciu o przebiegi czasowe, na którym Δt_k oznacza czas pomiędzy transmisją dwóch kolejnych sygnałów lokalizacyjnych i zakłada się, że jest on stały dla każdej SB, $\Delta t_{1,i}$ (i = 2, … N) reprezentuje różnicę w synchronizacji i-tej stacji bazowej względem SB_1 oraz $t_{a,k}$ i $t_{a,k+1}$ są czasami akwizycji danych pomiarowych w SR w momentach obserwacji k i k+1 i ich wartości maksymalne nie przekraczają Δt_k.



Rys. 2. Przebiegi czasowe w rozważanym problemie

Zatem, przy powyższych założeniach, na podstawie pomiarów jedynie czasów $t_{1i,k}$ (i = 2, … N) oraz $t_{1i,k+1}$ (i = 1, … N) będąca w ruchu SR jest w stanie dokonać estymacji swojego położenia.

3. ALGORYTM GENETYCZNY

Rozwiązanie układu równań nieliniowych (1) jest zagadnieniem złożonym. Zastosowanie klasycznych algorytmów nie daje zadowalających rezultatów. Zdecydowano się więc na algorytm genetyczny GA (*Genetic Algorithm*), który pomimo stosunkowo dużej złożoności obliczeniowej umożliwia znalezienie właściwego rozwiązania dla zadanych warunków początkowych. Warto zwrócić uwagę, że do rozwiązania układu równań (1) zastosowano tylko jeden algorytm, ponieważ głównym celem tego artykułu jest przedstawienie nowej metody lokalizacji, a nie optymalizacja metod rozwiązywania tego układu równań nieliniowych.

Zaproponowany algorytm genetyczny, który został zaimplementowany dla potrzeb przeprowadzonych badań symulacyjnych jest opisany poniżej i zaczerpnięty z [5]:

Step 0. Given W numerous $\mathbf{x}_0 \in \Re^n$ (initial population of chromosomes);
Set the threshold calculations *thold*;
Set the number of iteration *n_iter*;
Compute objective function $\mathbf{f}(\mathbf{x}_0)$ and $\|\mathbf{f}(\mathbf{x}_0)\|$;
Sort $\mathbf{x}_0$ by values $\|\mathbf{f}(\mathbf{x}_0)\|$;
If the lowest $\|\mathbf{f}(\mathbf{x}_0)\| \le thold$, then STOP, otherwise, set i:=1.

Step 1. Creation of a new population $\mathbf{x}_i$ for ½ W less matched population using ½ W better matched population (process of selection and mutation of chromosomes).

Step 2. Compute $\mathbf{f}(\mathbf{x}_i)$ and $\|\mathbf{f}(\mathbf{x}_i)\|$ for W numerous population of chromosomes;
Sort $\mathbf{x}_0$ by values $\|\mathbf{f}(\mathbf{x}_i)\|$;
If the lowest $\|\mathbf{f}(\mathbf{x}_i)\| \le thold$, then STOP.

Step 3. $i:=i+1$;
If $i \ge n_iter$, then STOP, otherwise go to **Step 1.**

W następnej części artykułu zostały przedstawione wyniki badań symulacyjnych, w których do rozwiązywania układu równań nieliniowych (1) zastosowano powyższy algorytm.

4. ANALIZA PARAMETRU PDOP

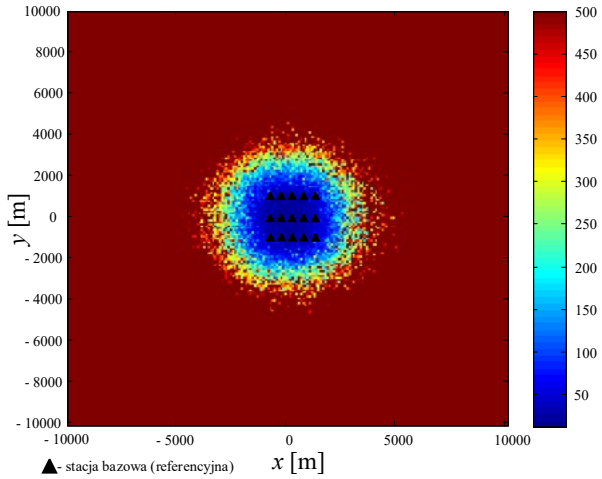
Dokładność estymacji położenia zależy od rozkładu geometrycznego stacji referencyjnych względem lokalizowanego obiektu, który często charakteryzowany jest parametrem PDoP (*Positional Dilution of Precision*) [6]. Im wyższa wartość tego parametru, tym niższa dokładność estymacji położenia obiektu uzyskiwana w systemie radiolokalizacyjnych pomimo utrzymania stałego poziomu dokładności pomiarów parametrów sygnałów radiowych. Współczynnik PDoP można wyznaczyć z zależności (4) [7]

$$P_D = \sqrt{tr[(\mathbf{J}^T\mathbf{J})^{-1}]}, \tag{4}$$

gdzie **J** reprezentuje macierz Jakobiego wyznaczoną na podstawie (1), natomiast $tr[\bullet]$ jest tzw. śladem macierzy. Stosując składnię uniwersalnego środowiska obliczeń matematycznych MATLAB oraz przyjęte powyżej oznaczenia dla zadanej liczby stacji bazowych (`l_ref`) macierz Jakobiego $[J_{i,j}]$ można utworzyć za pomocą następującego fragmentu pseudokodu:

```
for i=1:l_ref,
 J_i,1=(X_i-x_1,k-t_1i,k*v*V_x)/r_i,k-...
       (X_i-x_1,k-t_1i,k+1*v*V_x)/r_i,k+1;
 J_i,2=(Y_i-y_1,k-t_1i,k*v*V_y)/r_i,k-...
       (Y_i-y_1,k-t_1i,k+1*v*V_y)/r_i,k+1;
 J_i,3=(Z_i-z_1,k-t_1i,k*v*V_z)/r_i,k-...
       (Z_i-z_1,k-t_1i,k+1*v*V_z)/r_i,k+1;
 J_i,4=(X_i-x_1,k-t_1i,k*v*V_x)*t_1i,k/r_i,k-...
       (X_i-x_1,k-t_1i,k+1*v*V_x)*t_1i,k+1/r_i,k+1;
 J_i,5=(Y_i-y_1,k-t_1i,k*v*V_y)*t_1i,k/r_i,k-...
       (Y_i-y_1,k-t_1i,k+1*v*V_y)*t_1i,k+1/r_i,k+1;
 J_i,6=(Z_i-z_1,k-t_1i,k*v*V_z)*t_1i,k/r_i,k-...
       (Z_i-z_1,k-t_1i,k+1*v*V_z)*t_1i,k+1/r_i,k+1;
 J_i,7=((Z_i-z_1,k-t_1i,k*v*V_z)*t_1i,k*V_z+...
        (Y_i-y_1,k-t_1i,k*v*V_y)*t_1i,k*V_y+...
        (X_i-x_1,k-t_1i,k*v*V_x)*t_1i,k*V_x))/r_i,k-...
       ((Z_i-z_1,k-t_1i,k+1*v*V_z)*t_1i,k+1*V_z+...
        (Y_i-y_1,k-t_1i,k+1*v*V_y)*t_1i,k+1*V_y+...
        (X_i-x_1,k-t_1i,k+1*v*V_x)*t_1i,k+1*V_x))/r_i,k+1;
end
```

Do analizy rozkładu wartości współczynników PDoP dla opracowanej metody przyjęto następujący model: liczba oraz orientacja turbin wiatrowych jak na rys. 1, wysokość wieży pojedynczej turbiny $H = 100$ m, średnica rotora turbiny $D = 100$ m, odległości pomiędzy sąsiednimi turbinami wzdłuż osi x równają się $5{\cdot}D$, natomiast $10{\cdot}D$ wzdłuż osi y. Współczynnik PDoP wyznaczano na płaszczyźnie w kształcie kwadratu o boku $2{\cdot}R$ umieszczonej symetrycznie względem środka układu współrzędnych i zadanej wysokości z nad farmą wiatrową. Dla tego przypadku (`l_ref` = 15) macierz Jakobiego składa się z 15 wierszy i 7 kolumn. Przykładowe wyniki obliczeń numerycznych dla $R = 10\,000$ m i $z = 1\,000$ m przy losowo wybieranej orientacji trajektorii i prędkości ruchu przedstawiono na rys. 3.



Rys. 3. Przykładowy rozkład wartości parametru PDoP dla R = 10 000 m i z = 1 000 m

Na części obszaru w pobliżu stacji referencyjnych współczynnik PDoP dla rozważanej metody zmienia się od kilku do kilkudziesięciu. Tam też możemy się spodziewać stosunkowo niewielkich błędów estymacji położenia obiektów. Znacznie wyższe wartości PDoP, dochodzące do kilkuset, obserwowane są na krańcach analizowanego obszaru.

5. WYNIKI BADAŃ SYMULACYJNYCH

Badania symulacyjne efektywności pracy zaproponowanej metody przeprowadzono z zastosowaniem przyjętego na rys. 1 modelu sieci radiowej.

Oprogramowanie symulacyjne zostało wykonane w środowisku MATLAB. Podczas badań symulacyjnych uwzględniono błędy pomiarów czasów odbioru przez SR sygnałów lokalizacyjnych ($\Delta t_{1i,k}$ oraz $\Delta t_{1i,k+1}$). Do modelowania tych błędów wykorzystano funkcję *randn* w następujący sposób

$$\Delta t_{1i,k} = \varepsilon_t \cdot randn \quad \text{lub} \quad \Delta t_{1i,k+1} = \varepsilon_t \cdot randn, \tag{5}$$

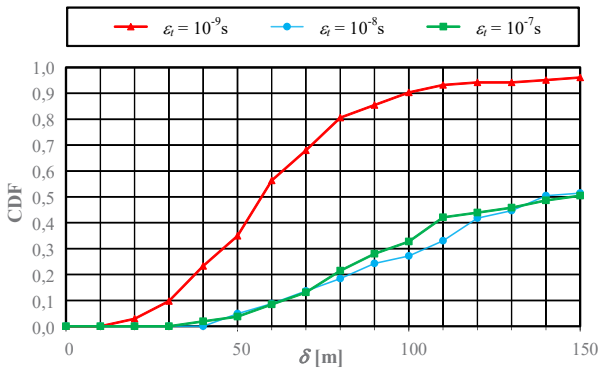
gdzie $\varepsilon_t = 10^{-9}$ s lub $\varepsilon_t = 10^{-8}$ s lub $\varepsilon_t = 10^{-7}$ s. W procesie estymacji położenia SR wykorzystano algorytm genetyczny opisany w sekcji 3 z następującymi parametrami: $W = 1\,000$, $thold = 0{,}01$ oraz $n_iter = 5\,000$. Przyjęto również, że samolot poruszał się ze średnią prędkością 250 m/s (900 km/h), natomiast czas pomiędzy nadawanymi przez poszczególne stacje bazowe sygnałów lokalizacyjnych był stały i wynosił 1 s. Ponadto, podczas symulacji, wewnątrz walca o promieniu $R = 10\,000$ m i wysokości $L = 10\,000$ m generowano losowo współrzędne początkowe SR, następnie na losowo wybieranym kierunku ruchu obliczano kolejne współrzędne SR zgodnie z wyżej przyjętymi założeniami. Za pomocą algorytmu genetycznego (GA) estymowano współrzędne

początkowe SR, wektory orientacji oraz prędkość poruszania się stacji ruchomej. W celu zmniejszenia złożoności obliczeniowej badań symulacyjnych w procesie rozwiązywania układu równań (1) dla każdego rozpatrywanego przypadku przyjęto wektor początkowy znajdujący się wewnątrz kuli o promieniu 500 m wokół rzeczywistego położenia SR. W badaniach symulacyjnych każdy przypadek powtórzono 10 000 razy. Warto w tym miejscu wspomnieć, że zwiększenie liczby powtórzeń powyżej tej wartości praktycznie nie poprawiało wiarygodności uzyskanych wyników, a jedynie dodatkowo wydłużało czas trwania badań symulacyjnych.

Na podstawie przeprowadzonych symulacji wyznaczono estymatę dystrybuanty błędów CDF (*Cumulative probability Distribution Function*) estymacji położenia samolotu δ, przy czym błąd ten określono zależnością (6)

$$\delta = \sqrt{\left(\hat{x}_{1,k} - x_{1,k}\right)^2 + \left(\hat{y}_{1,k} - y_{1,k}\right)^2 + \left(\hat{z}_{1,k} - z_{1,k}\right)^2} \quad (6)$$

gdzie $(x_{1,k}, y_{1,k}, z_{1,k})$ reprezentują rzeczywiste współrzędne początkowe stacji ruchomej, natomiast $(\hat{x}_{1,k}, \hat{y}_{1,k}, \hat{z}_{1,k})$ jego estymaty. Wyniki symulacji zostały przedstawione na rys. 4.



Rys. 4. Estymata dystrybuanty CDF błędów określania położenia samolotu

Ze wstępnej analizy rezultatów uzyskanych z badań symulacyjnych wynika duży potencjał zaproponowanego rozwiązania do lokalizacji obiektów latających w przestrzeni trójwymiarowej. Z pewnością, na dokładność estymacji położenia SR wpływa stabilność zastosowanego w stacjach referencyjnych wzorca częstotliwości. Stabilność na poziomie nanosekund umożliwia w 90 % przypadków uzyskiwanie dokładności nie gorszej niż 100 m na stosunkowo dużym obszarze działania systemu (walec o promieniu i wysokości 10 km). Dokładność ta jest porównywalna z rozmiarami samolotów i przynajmniej kilkanaście razy lepsza od stosowanej minimalnej separacji między samolotami w kontrolowanej przestrzeni powietrznej, zatem można ją uznać za zadowalającą w aplikacjach wspomagających nawigację lotniczą.

Praktyczne zastosowanie opracowanej metody wymaga w pierwszej kolejności opracowania efektywnego algorytmu rozwiązywania układu równań nieliniowych (1). Na potrzeby niniejszego referatu zdecydowano się zbadać jedynie algorytm genetyczny w prostej strukturze implementacyjnej w celu wykazania, że istnieje rozwiązanie tego układu równań. W ostatnich latach dynamicznie rozwijane są również algorytmy głębokiego uczenia do rozwiązywania złożonych problemów matematycznych. Jest to z pewnością obszar do dalszych prac autorów w tym zakresie.

6. PODSUMOWANIE

W artykule przedstawiono nową metodę estymacji położenia obiektów ruchomych, przeznaczoną dla asynchronicznych sieci radiowych utworzonych na bazie zestawu niezsynchronizowanych nadajników rozmieszczonych, na potrzeby analizy, na turbinach farm elektrowni wiatrowych. Na podstawie opracowanego oprogramowania symulacyjnego przeprowadzono badania efektywności pracy nowej metody, a wyniki można uznać za optymistyczne.

Do estymacji położenia obiektów nowa metoda potrzebuje przynajmniej 7 stacji referencyjnych. Z badań symulacyjnych wynika, że metoda ta zapewnia stosunkowo dużą dokładność wyznaczania położenia obiektów ruchomych w przestrzeni trójwymiarowej szczególnie dla wysokostabilnych wzorców częstotliwości (na poziomie pojedynczych nanosekund) stosowanych do budowy stacji bazowych. Zaproponowana metoda może stanowić rozwiązanie wspomagające klasyczne systemy nawigacji lotniczej oraz być pomocna podczas kontroli lotu statków powietrznych UAV (*Unmanned Aerial Vehicles*) – popularnych dronów, zarówno w obszarze lądowym jak i morskim ze szczególnym uwzględnieniem realizacji zadań inspekcyjnych w obrębie i sąsiedztwie farm wiatrowych.

LITERATURA

[1] Garbala Krzysztof, Tokajuk Aniela Marlena, Kałłaur Kamil, Cybulko Piotr. 2020. Klasyfikacja i charakterystyka turbin wiatrowych (silników wiatrowych). *Aparatura Badawcza i Dydaktyczna*, 25: 56–60.

[2] Stefański Jacek. 2018. Asynchroniczne metody radiolokalizacyjne. Ubiquitous Positioning, Wydawnictwo Naukowe PWN, Warszawa.

[3] Meyers Johan, Meneveau Charles. 2012. „Optimal Turbine Spacing in Fully Developed Wind-Farm Boundary Layers". *Wind Energy*, 15: 305-317

[4] Stefański Jacek, Sadowski Jarosław. 2018. „TDOA versus ATDOA for Wide Area Multilateration System". *EURASIP Journal on Wireless Communications and Networking*, (1): 1-13.

[5] Stefański Jacek, Sadowski Jarosław. 2016. „Analiza porównawcza metody TDOA i ATDOA w systemie WAM". *Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne*, 6: 297-300.

[6] Bard John D., Ham Fredric M. 1999. „Time Difference of Arrival Dilution of Precision and Applications". *IEEE Transactions on Signal Processing*, 47 (2): 521-523.

[7] Tsui James Bao-Yen. 2000, *Fundamentals of Global Positioning System Receivers. A Software Approach*. Wiley & Sons, New York.

ZASTOSOWANIE KRIGINGU MAP RSSI DO LOKALIZACJI WEWNĄTRZBUDYNKOWEJ METODĄ ODCISKÓW PALCÓW WSPOMAGANEJ FILTRACJĄ CZĄSTECZKOWĄ

APPLICATION OF KRIGING TO RSSI MAPS FOR INDOOR LOCATION WITH FINGERPRINTING ASSISTED BY PARTICLE FILTER

Robert Kawecki[1]; Sławomir Hausman[2]; Piotr Korbel[3];

[1] Robert Kawecki, Łódź, robert.kawecki@p.lodz.pl
[2] Sławomir Hausman, Łódź, slawomir.hausman@p.lodz.pl
[3] Piotr Korbel, piotr.korbel@p.lodz.pl

DOI: 10.15199/59.2023.4.52

Streszczenie: **W artykule zaprezentowano możliwość wykorzystania alternatywnej metody interpolacji rozkładów propagacji sygnału radiowego – kriging. Prezentowane referencyjne mapy RSSI są wykorzystywane jako zbiory danych do systemu lokalizacji w budynkach. System wykorzystuje metodę odcisków palców wspomaganą algorytmem filtracji cząsteczkowej. W ramach prezentowanych prac badawczych zweryfikowano, że możliwe jest efektywne wykorzystanie metody kriging do interpolacji danych pomiarowych RSSI i uzyskanie mniejszego błędu średniego i medianowego niż wykorzystanie danych pomiarowych lub rozkładów RSSI wygenerowanych z wykorzystaniem interpolacji liniowej.**

Abstract: **The article presents the possibility of using an alternative method of interpolation of radio signal propagation distribution sets – kriging. The presented reference RSSI distributions are used as data sets for the location system in buildings. The localization system uses the fingerprint method supported by a molecular filtration algorithm. As part of the presented research, it was verified that it is possible to effectively use the kriging method to interpolate RSSI measurement data and obtain a smaller average and median error than using measurement data or RSSI distributions generated using linear interpolation.**

Słowa kluczowe: **kriging, RSSI, lokalizacja wewnątrz budynków, metoda odcisków palców, filtracja cząsteczkowa, Bluetooth, BLE**

Keywords: **kriging, RSSI, indoor localization, fingerprinting, particle filter, Bluetooth, BLE**

1. WSTĘP

Lokalizowanie osób, obiektów i urządzeń (w tym autonomicznych pojazdów i robotów samojezdnych) w pomieszczeniach jest postrzegane jako kluczowa technologia umożliwiająca realizację wielu zastosowań w obszarze tzw. inteligentnego środowiska [1]. Jest to też ważny element tzw. czwartej rewolucji przemysłowej [2], bez którego nie mogłyby funkcjonować mobilne urządzenia autonomiczne. Duże przestrzenie, takie jak wielkopowierzchniowe obiekty mieszkaniowe, hotele, centra handlowe, hale produkcyjne, biurowce, lotniska i budynki administracyjne, mogą być trudne w nawigacji. Istniejące systemy radionawigacji satelitarnej nie mogą być skutecznie stosowane wewnątrz budynków ze względu na tłumienie sygnału przez ich elementy konstrukcyjne i propagację wielodrogową [3].

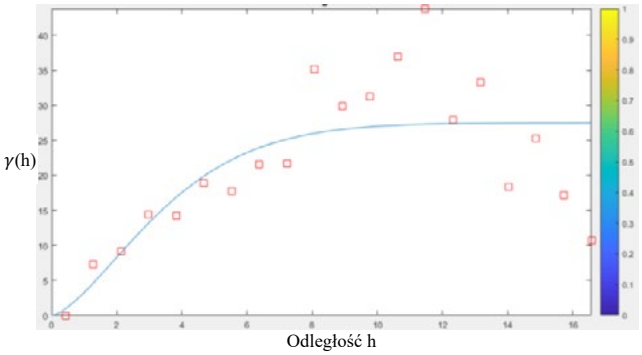
W ostatnich dziesięcioleciach wzrosło zainteresowanie radiowymi technikami lokalizacji wewnątrz budynków [4]. Niektórzy autorzy skupili się na źródłach danych jakie są rejestrowane w środowisku budynkowym i wykorzystywane do lokalizacji w budynkach [3]. Pozycjonowanie w pomieszczeniach wymaga zwykle wykorzystania sygnałów z wielu źródeł, np. nadajników Wi-Fi [5] lub Bluetooth [6], czujników inercyjnych [3], stacji bazowych sieci komórkowych [7]. Kluczowe znaczenie dla działania systemów lokalizacji mają algorytmy wyznaczania położenia. W prezentowanych badaniach wykorzystano opracowany i zaimplementowany w ramach badań zmodyfikowany algorytm odcisków palców (ang. fingerprinting, FP) wspomagany filtracją cząsteczkową [8]. Polega on na ocenie podobieństwa obserwowanych w danej lokalizacji parametrów sygnałów radiowych z danymi zapisanymi w bazie danych, tzw. mapie referencyjnej. Wykorzystanie tej metody wymaga tym samym przygotowania baz danych referencyjnych opisujących charakterystykę przestrzenną sygnałów radiowych, np. rozkładów RSSI z dostępnych nadajników w obszarze zainteresowania.

W literaturze prezentowane są dwie najczęściej popularne metody tworzenia baz danych o obiekcie: przez gromadzenie i przetwarzanie informacji przez użytkowników poruszających się w budynku [9], lub wykorzystywanie informacji uzyskiwanych z istniejących źródeł, np. poprzez wykonywanie pomiarów mocy odebranej z nadajników radiowych, zlokalizowanych w budynku [4]. Takie pomiary są zazwyczaj bardzo pracochłonne i czasochłonne. W praktyce wykonanie pomiarów w budynku, zwłaszcza w gęstej siatce punktów pomiarowych jest praktycznie niemożliwe – w związku z tym, w celu przygotowania ich do możliwości wykorzystania w systemie lokalizacyjnym poddawane są interpolacji.

W ramach prowadzonych badań zweryfikowana została alternatywna metoda interpolowania referencyjnych map RSSI dla celów lokalizacyjnych. W tym rozdziale omówiona została metoda krigingu, dzięki której możliwe jest przygotowanie mapy rozkładu RSSI wykorzystując mniejszą liczbę wykonanych pomiarów w budynku niż w przypadku użycia obecnie bardzo często stosowanej metody interpolacji liniowej.

2. INTERPOLACJA METODĄ KRIGINGU

Kriging to statystyczna metoda estymacji wartości funkcji za pomocą liniowej kombinacji znanych wartości (próbek) tej funkcji w skończonej liczbie punktów. Danym próbkom (punktom pomiarowym) znajdującym się wewnątrz obszaru estymacji (obszaru wyszukiwania próbek) przypisuje się odpowiednie wagi zwane współczynnikami krigingu w taki sposób, aby zminimalizować średniokwadratowy błąd estymacji (wariancję krigingu).



Rys. 1 Przykładowy wykres semiwariogramu wygenerowany przy tworzeniu mapy RSSI metodą krigingu

Semiwariogram to podstawowe narzędzie służące do estymacji i badania struktury zmienności badanych zjawisk w geostatystyce. Definiuje się go jako połowa średniej kwadratowej różnicy między dwiema wartościami cechy mierzalnej, pomiędzy którymi odległość równa jest w przybliżeniu wektorowi h. Rysunek 1 przedstawia przykładowy wykres semiwariogramu. Wyliczone wartości funkcji γ(h) aproksymuje się modelami teoretycznymi (funkcjami analitycznymi), które następnie wykorzystuje do ustalenia wartości współczynników wagowych. W tej metodzie przyjmuje się następujące założenia: dane są izotropowe (wykazują te same cechy we wszystkich kierunkach poszukiwania), średnie lokalne nie mają ścisłego związku ze średnią z całej populacji próbek, a podczas oceny średniej wybiera się jedynie punkty pochodzące z lokalnego sąsiedztwa.
W badaniach wykorzystano tzw. kriging „zwyczajny" (ang. ordinary kriging). W krigingu wartość szacowana (interpolowana) ma postać średniej ważonej:

$$z_k^x = \sum_{i=1}^{n} w_i \cdot z_i \qquad (1)$$

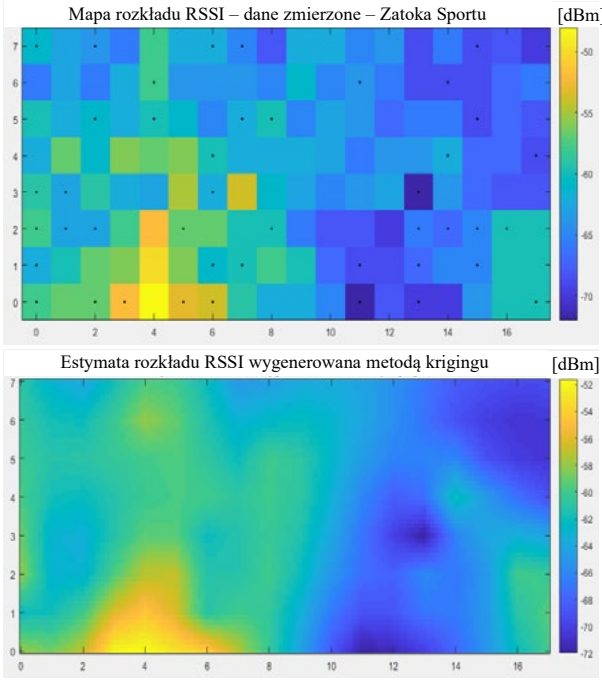
gdzie:
w_i – współczynnik wagowy przypisany pojedynczej obserwacji,
z_i – wartość badanego parametru w pojedynczym punkcie pomiarowym,
n – liczba danych uwzględnianych w ocenie wartości parametru.
Wartości współczynników wagowych w_i przypisywanych poszczególnym obserwacjom przy szacowaniu parametrów w węzłach zadanej sieci interpolowanej obliczana jest w za pomocą układu równań krigingu na podstawie znajomości struktury zmienności badanego parametru.
Informacji na temat struktury zmienności parametru dostarcza semiwariogram opisany formułą:

$$\gamma(h) = \frac{1}{2n_h}\sum_{i=1}^{n_h}(z_{i+h} - z_i)^2 \qquad (2)$$

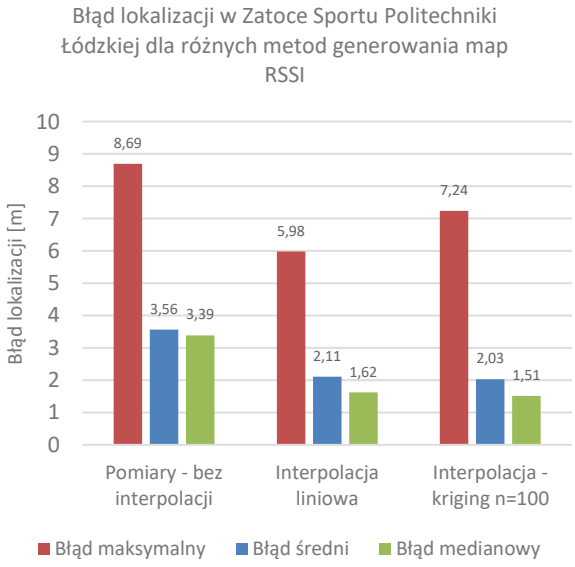
gdzie:
z_{z+h}, z_i – wartości badanego parametru w punktach odległych o h,
n_h – liczba punktów pomiarowych odległych o h.
Semiwariogramy przedstawiają zróżnicowanie wartości parametrów w zależności od odległości między punktami pomiarowymi, a zatem strukturę ich zmienności, w sposób pośredni charakteryzując autokorelację między obserwacjami. Wyliczone wartości funkcji $\gamma(h)$ aproksymuje się modelami teoretycznymi (funkcjami analitycznymi), które następnie wykorzystuje się do ustalenia wartości współczynników wagowych.
Metoda krigingu była już wykorzystywana do interpolacji zmierzonych wartości RSSI w budynkach w celu utworzenia referencyjnych map rozkładu sygnału radiowego do lokalizacji w budynkach [10], [11]. Celem badań było porównanie błędu lokalizacji w budynkach z wykorzystaniem map rozkładu RSSI wygenerowanych metodą krigingu, interpolacji liniowej danych pomiarowych oraz wygenerowanych metodą symulacji komputerowych.

3. EKSPERYMENT BADAWCZY

Celem omawianych tu badań była weryfikacja możliwości wykorzystania metody krigingu do tworzenia map rozkładu RSSI w budynku w zastosowaniu do lokalizacji. Metoda umożliwia utworzenie interpolowanej mapy rozkładu danej wartości bazując na stosunkowo niedużej liczbie pomiarów wykonanych w środowisku rzeczywistym. W tym celu wykorzystano dane pomiarowe zebrane w jednym ze środowisk wewnątrzbudynkowych – w budynku obiektu „Zatoka Sportu" Politechniki Łódzkiej (Dane zebrane podczas przeprowadzonej kampanii pomiarowej zostały udostępnione na serwerze github: https://github.com/robertkawecki7/IndoorPositioningDatasets). Zbiór zawiera wartości mocy odebranej zmierzone w węzłach siatki o rozmiarach 8 x 18 i o oczku 2 m. Metoda krigingu zakłada wybranie n losowych próbek danych źródłowych do przeprowadzenia procesu interpolacji. W kolejnym etapie na podstawie wybranych próbek i nadanych im wag, wykonywana jest procedura krigingu w celu wyznaczenia zbioru interpolowanego. Na rysunku 2 przedstawiono rozkład przestrzenny RSSI otrzymany za pomocą pomiarów w budynku oraz poniżej, interpolowany metodą krigingu. Interpolację wykonuje się raz, na etapie przygotowania map referencyjnych RSSI w danym budynku. W związku z tym przetwarzanie danych RSSI może być realizowane w trybie offline (czyli bez wykorzystania urządzeń mobilnych), co powoduje, że złożoność obliczeniowa algorytmu interpolacji nie jest w przypadku tego typu systemów ograniczona do mocy obliczeniowych urządzeń mobilnych, gdzie estymowana jest lokalizacji użytkownika.
Po wygenerowaniu zbioru danych stanowiących referencyjny rozkład RSSI w budynku wykorzystując opracowaną platformę symulacyjną [8] przeprowadzono symulację przemieszczania się użytkownika, w celu wyznaczenia estymat lokalizacji metodą odcisków palców i wyznaczenie błędu lokalizacji i porównanie z błędem lokalizacji dla innych metod tworzenia map referencyjnych RSSI. Na rysunku 3 pokazano porównanie otrzymanych błędów lokalizacji dla trzech różnych metod generowania referencyjnych rozkładów RSS.



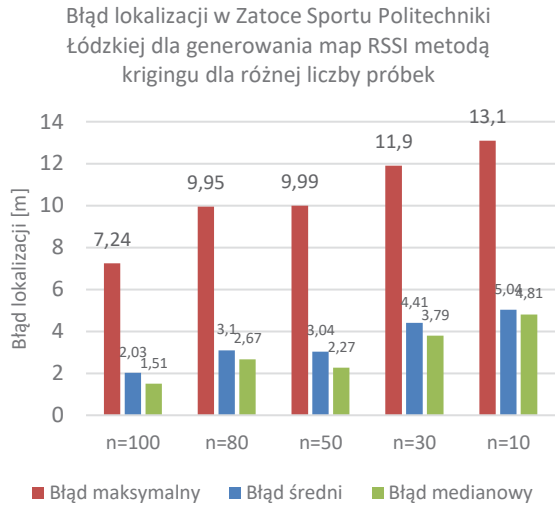
Rys. 2 Rozkłady przestrzenne RSSI otrzymane za pomocą pomiarów w środowisku wewnątrzbudynkowym oraz interpolowane metodą kriginu.



Rys. 3 Błąd lokalizacji w Zatoce Sportu Politechniki Łódzkiej dla różnych metod generowania map RSSI (za pomocą wykonywania pomiarów, bez interpolacji, z wykorzystaniem interpolacji liniowej oraz interpolacji metodą kriginu).

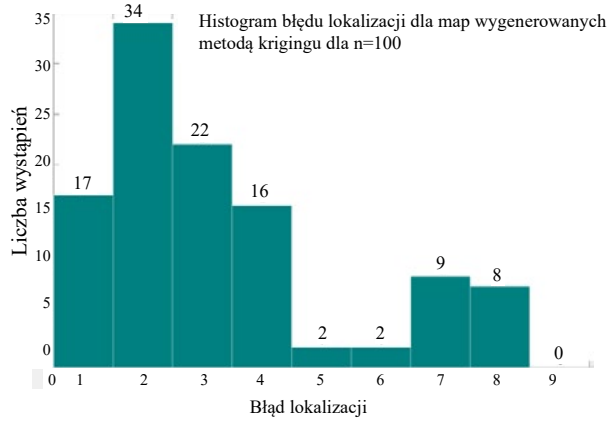
4. WPŁYW LICZBY PRÓBEK KRIGINGU NA BŁĄD LOKALIZACJI

Przeprowadzono także weryfikację wpływu liczby wykorzystanych próbek z bazy danych pomiarowych wziętych pod uwagę do interpolacji za pomocą metody kriginu. Na rysunku 4 przedstawiono podsumowanie przeprowadzonej analizy w formie wykresów słupkowych. Można zauważyć, że im większa liczba wykorzystanych próbek tym możliwe jest otrzymanie mniejszego błędu maksymalnego, średniego, a także mediany błędu lokalizacji. Zmniejszenie liczby próbek wykorzystanych do kriginu wpływa na wzrost wartości błędu maksymalnego. Dla rozpatrywanego scenariusza weryfikacyjnego zmniejszenie liczby próbek ze 100 do 80 powoduje zwiększenie błędu maksymalnego o około 27%. Dla mediany błędu – widoczne jest znaczne pogorszenie, które można uznać za nieakceptowalne dla tej klasy metod występuje dopiero po zmniejszeniu liczby próbek o połowę (ze 100 do 50).



Rys. 4 Błąd lokalizacji w Zatoce Sportu Politechniki Łódzkiej dla różnej liczby próbek n uwzględnionej do wygenerowania map RSSI metodą kriginu.

Na rysunku 5 zaprezentowano histogram rozkładu błędu lokalizacji otrzymanego dla jednego z prezentowanych scenariuszy weryfikacyjnych (dane wygenerowane metodą kriginu, dla n=100, zasięg=12).



Rys. 5 Histogram rozkładu błędu lokalizacyjnego otrzymanego dla jednego z prezentowanych scenariuszy weryfikacyjnych (dane wygenerowane metodą kriginu, dla n=100, zasięg=12).

Około 81% z uzyskanych błędów lokalizacji mieści się w granicy 0 – 4 m (dla rozpatrywanego przypadku mediana błędu wyniósł 1,51 m). Dane zaprezentowane na

wykresach na rysunku 3 wskazują, że wykorzystanie danych wygenerowanych metodą krigingu skutkuje zmniejszeniem mediany błędu lokalizacyjnego względem danych otrzymanych metodą interpolacji liniowej o 7%, a błędu średniego o 4%. W porównaniu do wykorzystania danych pomiarowych jako źródła dla algorytmu odcisków palców, bez uwzględnienia interpolacji otrzymano mniejszy błąd średni o 43%, medianę błędu o 55%. Zastosowanie metody krigingu do tworzenia mapy referencyjnej RSSI skutkuje zmniejszeniem błędu maksymalnego lokalizacji względem danych pomiarowych bez uwzględnienia interpolacji o 17% oraz zwiększenie błędu maksymalnego względem interpolacji liniowej o 21%.

5. PODSUMOWANIE

W badaniach prezentowanych w niniejszym artykule zweryfikowano alternatywną metodę interpolacji danych pomiarów RSSI w budynku – kriging. Przedstawiono wpływ dwóch podstawowych parametrów metody – czyli liczba próbek branych do interpolacji oraz zasięg, czyli odległość między znanymi próbkami, a próbką, która jest interpolowana. Przeanalizowano możliwość wykorzystania referencyjnych rozkładów RSSI interpolowanych metodą krigingu w autorskim systemie lokalizacyjnym wykorzystującym metodę odcisków palców. Zmierzono błąd lokalizacyjny za pomocą zaimplementowanego algorytmu odcisków palców wykorzystującego rozkłady RSSI wygenerowane metodą krigingu.

Podsumowując – wykorzystanie metody krigingu stanowi wartościową alternatywę dla metod interpolacji podczas generowania rozkładów referencyjnych RSSI dla celów lokalizacyjnych przy stosunkowo niewielkiej liczbie zmierzonych danych w budynku. Dzięki wykorzystaniu metody krigingu istnieje możliwość wykorzystania danych pomiarowych do wygenerowania referencyjnej mapy RSSI dla algorytmu odcisków palców nawet w przypadku, gdy liczba punktów pomiarowych jest niewielka. Analiza porównawcza błędów lokalizacyjnych dla map RSSI wygenerowanych metodami interpolacji liniowej oraz interpolacji metodą krigingu potwierdziły dobrą dokładność lokalizacji (a nawet zmniejszenie błędu), co pozwala na stwierdzenie, że wykorzystanie metody kriging do interpolacji danych pomiarowych (zwłaszcza jeśli jest ich nieduża liczba, lub pomiary wykonane są w nieregularnych odstępach) umożliwia efektywne lokalizowanie osób i przedmiotów w budynkach z wykorzystaniem zaproponowanego algorytmu odcisków palców wspomaganego filtracją cząsteczkową.

LITERATURA

[1] I. Ashraf, S. Hur, and Y. Park, “Smartphone sensor based indoor positioning: Current status, opportunities, and future challenges,” *Electron.*, vol. 9, no. 6, 2020, doi: 10.3390/electronics9060891.

[2] “Raport z badania ‘Czwarta rewolucja przemysłowa i jej wpływ na rynek pracy’ przeprowadzonego przez Instytut Analiz Rynku Pracy Sp. z o.o. w 2020 roku na zlecenie Polskiej Agencji Rozwoju Przedsiębiorczości, Warszawa 2020 rok. http://www.parp.gov.pl,” 2020.

[3] F. Zafari, A. Gkelias, and K. K. Leung, “A Survey of Indoor Localization Systems and Technologies,” *IEEE Commun. Surv. Tutorials*, 2019, doi: 10.1109/COMST.2019.2911558.

[4] T. Yang, A. Cabani, and H. Chafouk, “A survey of recent indoor localization scenarios and methodologies,” *Sensors*, vol. 21, no. 23, 2021, doi: 10.3390/s21238086.

[5] S. Chan and G. Sohn, “Indoor Localization Using Wi-Fi Based Fingerprinting and Trilateration Techiques for Lbs Applications,” *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. XXXVIII-4/, pp. 1–5, 2012, doi: 10.5194/isprsarchives-xxxviii-4-c26-1-2012.

[6] Y. Wang, X. Yang, Y. Zhao, Y. Liu, and L. Cuthbert, “Bluetooth positioning using RSSI and triangulation methods,” *2013 IEEE 10th Consum. Commun. Netw. Conf. CCNC 2013*, pp. 837–842, 2013, doi: 10.1109/CCNC.2013.6488558.

[7] M. Driusso, C. Marshall, M. Sabathy, F. Knutti, H. Mathis, and F. Babich, “Indoor positioning using LTE signals,” *2016 Int. Conf. Indoor Position. Indoor Navig. IPIN 2016*, 2016, doi: 10.1109/IPIN.2016.7743656.

[8] R. Kawecki, S. Hausman, and P. Korbel, “Performance of Fingerprinting-Based Indoor Positioning with Measured and Simulated RSSI Reference Maps,” *Remote Sens.*, vol. 14, no. 9, 2022, doi: 10.3390/rs14091992.

[9] S. H. Jung and D. Han, “Automated Construction and Maintenance of Wi-Fi Radio Maps for Crowdsourcing-Based Indoor Positioning Systems,” *IEEE Access*, 2017, doi: 10.1109/ACCESS.2017.2780243.

[10] K. Kawamoto, Y. Tomura, and K. Okamoto, “Kriging-based prediction and interpolation for modeling pedestrian dynamics,” *IFSA-SCIS 2017 - Jt. 17th World Congr. Int. Fuzzy Syst. Assoc. 9th Int. Conf. Soft Comput. Intell. Syst.*, vol. 0, no. 4, pp. 4–7, 2017, doi: 10.1109/IFSA-SCIS.2017.8023264.

[11] W. C. M. Van Beers and J. P. C. Kleijnen, “Kriging interpolation in simulation: A survey,” *Proc. - Winter Simul. Conf.*, vol. 1, pp. 113–120, 2004, doi: 10.1109/wsc.2004.1371308.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

WSKAZYWANIE PUNKTU STARTOWEGO DLA ITERACYJNYCH ALGORYTMÓW OBLICZANIA POŁOŻENIA Z UŻYCIEM SIECI NEURONOWEJ

STARTING POINT SELECTION FOR ITERATIVE POSITION ESTIMATION ALGORITHMS USING NEURAL NETWORK

Jarosław Sadowski[1]; Jacek Stefański[2]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, jarsad@eti.pg.edu.pl
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, jstef@eti.pg.edu.pl

DOI: 10.15199/59.2023.4.53

Streszczenie: **W referacie przedstawiono wyniki badań nad możliwością wskazywania punktu startowego do pierwszej iteracji dla algorytmu iteracyjnego obliczania położenia w systemie lokalizacji dwuwymiarowej. Do wskazywania punktu startowego użyto jednokierunkowej sieci neuronowej a celem badań było znalezienie jak najmniejszej struktury sieci, pozwalającej na zbieżność algorytmu estymacji położenia w całym obszarze badań.**

Abstract: **The paper presents the results of a study on the possibility of starting point selection for the first iteration for an iterative position calculation algorithm in a two-dimensional location system. A feedforward neural network was used to indicate the starting point and the aim of the study was to find the smallest possible network structure, allowing the position estimation algorithm to converge over the entire study area**

Słowa kluczowe: **algorytm iteracyjny, jednokierunkowa sieć neuronowa, radiolokalizacja, radionawigacja**

Keywords: **feedforward neural network, iterative algorithm, radiolocalization, radionavigation**

1. WSTĘP

Określanie położenia terminali ruchomych w sieciach radiowych, czyli realizacja usługi lokalizacyjnej, wymaga zrealizowania pomiarów określonych parametrów sygnałów radiowych, wysyłanych lub odbieranych przez lokalizowany węzeł ruchomy, oraz przeliczenia wyników pomiarów na parametry lokalizacyjne, najczęściej będące estymatami współrzędnych terminala w zdefiniowanym układzie współrzędnych. Poza nielicznymi wyjątkami, zależności pomiędzy wynikami zrealizowanych pomiarów a poszukiwanymi współrzędnymi są nieliniowe, a w niektórych przypadkach nawet niejednoznaczne. Z tego względu opracowanych zostało wiele algorytmów obliczania położenia, zarówno bazujących na ogólnych algorytmach rozwiązywania układów równań nieliniowych jak i algorytmów dedykowanych ściśle do zastosowań lokalizacyjnych [1][2]. Te algorytmy obliczeniowe mogą być podzielone na dwie grupy. Grupa pierwsza to algorytmy nieiteracyjne, w których wynikowe współrzędne terminala są uzyskiwane po jednokrotnym wykonaniu zadanej sekwencji operacji matematycznych. Dla odmiany algorytmy z drugiej grupy to algorytmy iteracyjne, czyli takie, w których wymagane jest wielokrotne powtórzenie określonej sekwencji operacji, a z każdym powtórzeniem uzyskuje się coraz lepsze przybliżenie wyniku końcowego. Teoretycznie, dla uzyskania dokładnego wyniku wymagana może być nieskończona liczba iteracji, w praktyce jest ona zawsze ograniczona a obliczenia są kończone po uzyskaniu zadowalająco małego błędu równań pozycyjnych. Konieczność wielokrotnego powtarzania obliczeń jest uznawana za istotną wadę algorytmów iteracyjnych, przez którą ich złożoność obliczeniowa jest zwykle znacznie większa niż algorytmów nieiteracyjnych. Ponadto, do wad algorytmów iteracyjnych zaliczyć należy problem wyboru punktu początkowego do pierwszej iteracji. Rozpoczęcie obliczeń od niewłaściwie dobranego punktu startowego może skutkować zwiększeniem liczby iteracji, wymaganych do uzyskania rozwiązania końcowego, zbieżność algorytmu do niewłaściwego punktu odpowiadającego lokalnemu, a nie globalnemu minimum błędu równań pozycyjnych a nawet rozbieżność algorytmu.

Pomimo wyżej wymienionych wad, iteracyjne algorytmy obliczania położenia mogą być niezastąpione w pewnych zastosowaniach, np. w systemach pozycyjnych, w których powiązanie bezpośrednio mierzonej wielkości fizycznej, np. czasu, z obowiązującymi w systemie zależnościami geometrycznymi, np. odległościami lub różnicami odległości, jest nieliniowe i zależne od położenia terminala ruchomego, które to położenie ma być dopiero obliczone. Ponadto, algorytmy nieiteracyjne zwykle zakładają gaussowski rozkład błędów pomiarowych, podczas gdy w algorytmach iteracyjnych dość łatwo można zrealizować obsługę przypadków z niegaussowskim, a nawet niesymetrycznym rozkładem błędów. Biorąc pod uwagę te obszary zastosowań, w których algorytmy iteracyjne przewyższają inne metody obliczania położenia, postanowiono podjąć badania w zakresie wskazywania punktu startowego do pierwszej iteracji w celu zagwarantowania prawidłowego wskazywania położenia w całym obszarze działania systemu lokalizacyjnego.

Ponieważ istotny wpływ na zbieżność algorytmów iteracyjnych ma geometria układu stacji bazowych (węzłów referencyjnych), do badań przyjęto rzeczywiste rozmieszczenie czterech stacji bazowych LTE (*Long Term Evolution*) jednego z operatorów telefonii komórkowej, zainstalowanych w Gdańsku. Przyjęto także, że usługa lokalizacyjna jest realizowana w wariancie dwuwymiarowym (2D), na podstawie pomiarów różnic cza-

sów dotarcia sygnałów TDOA (*Time Difference Of Arrival*). Taki sposób estymacji położenia terminali komórkowych jest przewidziany w standardzie LTE i 5G, więc przyjęte założenia odpowiadają możliwym warunkom realizacji usługi lokalizacyjnej w rzeczywistej sieci komórkowej. Natomiast jako cel badań przyjęto znalezienie jak najprostszej struktury sieci neuronowej typu jednokierunkowego FNN (*Feedforward Neural Network*), która na bazie zrealizowanych pomiarów różnic odległości pozwoli na wskazanie punktu startowego do pierwszej iteracji w taki sposób, by zastosowany algorytm iteracyjnej obliczania położenia w każdym punkcie obszaru badań uzyskiwał zbieżność do właściwego rozwiązania.

2. ITERACYJNE OBLICZANIE POŁOŻENIA

Do badań problemu zbieżności wybrano stosunkowo prosty algorytm iteracyjny Gaussa-Newtona [1][2]. Dla określonej pozycji terminala ruchomego: (x, y), system TDOA pozwala na wykonanie pomiarów różnic odległości od par stacji bazowych o numerach m, n:

$$d_{m,n} = \sqrt{(x - X_m)^2 + (y - Y_m)^2} - \sqrt{(x - X_n)^2 + (y - Y_n)^2} \quad (1)$$

gdzie (X_m, Y_m), (X_n, Y_n) to współrzędne stacji bazowych odpowiednio o numerach m i n. Estymaty współrzędnych $(\hat{x}, \hat{y})$ w k+1-szym kroku są obliczane iteracyjnie na podstawie wartości z kroku k:

$$\begin{bmatrix} \hat{x} \\ \hat{y} \end{bmatrix}_{k+1} = \begin{bmatrix} \hat{x} \\ \hat{y} \end{bmatrix}_k + \boldsymbol{\delta}_k \quad (2)$$

według zależności:

$$\boldsymbol{\delta}_k = \left(\mathbf{A}_k^T \mathbf{R}^{-1} \mathbf{A}_k \right)^{-1} \cdot \mathbf{A}_k^T \mathbf{R}^{-1} \mathbf{z}_k \quad (3)$$

w której $\mathbf{A}_k$ jest macierzą pochodnych cząstkowych równań pozycyjnych w punkcie $(\hat{x}, \hat{y})$ w kroku k, $\mathbf{R}$ to macierz kowariancji błędów zaś z_k to wektor różnic między odległościami pomierzonymi i estymowanymi w k-tym kroku obliczeń [2].

Zasady realizacji obliczeń iteracyjnych w innych algorytmach, np. Levenberga-Marquardta są podobne, z różnicami przede wszystkim w postaci równania (3). Eksperymentalnie sprawdzono, że przedstawione dalej wnioski z badań wspomagania algorytmu Gaussa-Newtona przy użyciu sieci neuronowej pozostają ważne także dla innych algorytmów iteracyjnych.

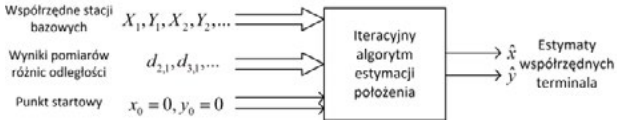
3. ZGRUBNA ESTYMACJA POZYCJI STARTOWEJ DO PIERWSZEJ ITERACJI

Na potrzeby prowadzonych badań, współrzędne czterech rzeczywistych stacji bazowych LTE z obszaru Gdańska zostały przedstawione w lokalnym, kartezjańskim układzie współrzędnych, którego początek (0, 0) został zdefiniowany w przybliżeniu w środku obszaru ograniczonego stacjami bazowymi. Cztery stacje to minimalna liczba, niezbędna dla jednoznacznego wskazania położenia w systemie TDOA przy lokalizacji dwuwymiarowej, gdyż przy mniejszej liczbie stacji występuje możliwość uzyskania dwóch punktów przecięcia wszystkich hiperbol pozycyjnych dający niejednoznaczność wskazania pozycji. Współrzędne poszczególnych stacji zebrano w tab. 1. Ponieważ jednak z praktyki wiadomo, że problem zbieżności algorytmów iteracyjnych może się nasilać poza obszarem ograniczonym zestawem używanych stacji bazowych, symulacje przeprowadzono dla większego obszaru, do +/- 700 m od początku układu współrzędnych.

Tab. 1. Przyjęte współrzędne stacji bazowych w lokalnym układzie odniesienia

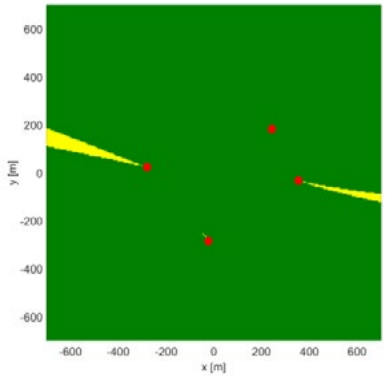
Numer stacji	*X [m]*	*Y [m]*
1	243	182
2	-23	-283
3	-278	23
4	352	-33

Jako scenariusz referencyjny przyjęto, że praca algorytmu iteracyjnej estymacji położenia jest rozpoczynana zawsze od punktu początkowego znajdującego się w początku układu współrzędnych, tj. x_0=0, y_0=0. Algorytm iteracyjny otrzymuje zatem jako dane wejściowe współrzędne wszystkich stacji bazowych i wyniki pomiarów różnic odległości, jak przedstawiono schematycznie na rys. 1.



Rys. 1. Schemat przepływu danych przy estymacji położenia z użyciem algorytmu iteracyjnego

Przy stałym punkcie startowym algorytm Gaussa-Newtona wykazał problemy ze zbieżnością przede wszystkim w dwóch obszarach położonych po przeciwnej stronie stacji bazowej nr 3 i 4 względem środka układu współrzędnych. W tych obszarach uzyskano współrzędne wynikowe $(\hat{x}, \hat{y})$ różniące się od pozycji rzeczywistej (x, y) o więcej niż 1m. Nie zaobserwowano przypadków rozbieżności algorytmu w badanym przypadku, choć nie można wykluczyć możliwości ich wystąpienia w innych scenariuszach. W całym badanym obszarze uzyskano zbieżność w 78071 z 78961 punktów, co stanowi 98,87%, przy czym średnia liczba iteracji algorytmu Gaussa-Newtona wyniosła 8,16 a maksymalna 71.

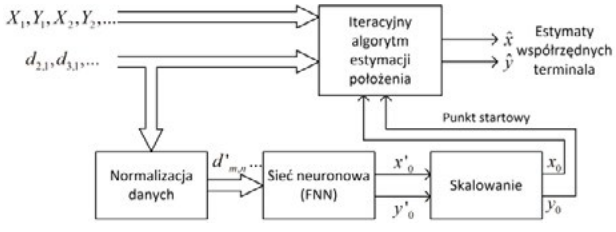


Rys. 2. Obszary zbieżności do nieprawidłowego wyniku, algorytm Gaussa-Newtona.

Zdefiniowanie punktu startowego do pierwszej iteracji w pobliżu rzeczywistej, poszukiwanej pozycji terminala ruchomego, znacząco redukuje ryzyko zbieżności do nieprawidłowych współrzędnych. Z tego powodu podjęto próbę zbadania, czy możliwe jest wskazywanie pozycji początkowej przy użyciu sieci neuronowej typu FNN [3]. Sieć będzie wyuczona do pracy wyłącznie ze stacjami bazowymi rozmieszczonymi zgodnie z tab. 1, więc jej danymi wejściowymi będą tylko wartości różnic odległości $d_{m,n}$, poddane normalizacji dla uzyskania zakresu zmienności wszystkich danych od -1 do +1:

$$d'_{m,n} = \frac{d_{m,n}}{\sqrt{\left(X_m - X_n\right)^2 + \left(Y_m - Y_n\right)^2}} \tag{4}$$

Również wyniki działania sieci, czyli estymaty współrzędnych (x_0, y_0) do pierwszej iteracji, zostały znormalizowane do wartości z tego samego zakresu. Schemat blokowy przepływu danych w przypadku estymacji punktu początkowego siecią neuronową przedstawiono na rys. 3.

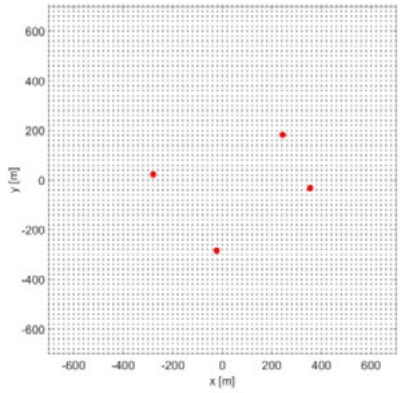


Rys. 3. Sposób użycia sieci FNN do wskazywania punktu startowego do pierwszej iteracji

Efektywność realizacji zadań przez sieci neuronowe jest zależna nie tylko od ich struktury, tj. liczby warstw, typu warstw, obecności pamięci, rodzaju funkcji aktywacji, połączeń między warstwami czy sprzężeń zwrotnych, ale też od doboru danych treningowych i realizacji procesu uczenia sieci [3]. W przypadku sformułowanego problemu: wskazywanie punktu startowego w taki sposób, by uzyskać zbieżność iteracyjnej estymacji położenia w całym analizowanym obszarze, definiowanie zbioru danych wzorcowych do trenowania sieci nie jest zadaniem oczywistym, gdyż zbieżność dla określonego zestawu wyników pomiarów różnic odległości można uzyskać w ogólności dla nieskończenie dużej liczby różnych współrzędnych punktu startowego, przy czym nie każdy zestaw współrzędnych pozwoli na uzyskanie wyniku końcowego po takiej samej liczbie iteracji. Na potrzeby badań przyjęto, że sieć będzie uczona przy pomocy rzeczywistych współrzędnych punktów, dla których obliczono wartości różnic odległości $d_{m,n}$, przy czym rozważono przypadek regularnego i nieregularnego generowania danych treningowych.

3.1. Regularne rozmieszczenie punktów do trenowania sieci neuronowej

Pierwszym podejściem do problemu uczenia sieci neuronowej było założenie, że metodą do uzyskania zbieżności iteracyjnych algorytmów estymacji położenia jest zgrubne wskazanie pozycji terminala ruchomego przez sieć z jednakowymi oczekiwaniami co do dokładności estymacji współrzędnych punktu startowego w zdefiniowanym obszarze funkcjonowania systemu. W takiej sytuacji jako dane treningowe użyto punktów rozmieszczonych regularnie w siatce o kroku 20 m (rys. 4), uzyskując zbiór 5041 zestawów różnic odległości i odpowiadających im współrzędnych wzorcowych. Algorytm trenowania sieci dzielił ten zbiór losowo w proporcjach: 70% dane treningowe, 15% dane walidacyjne i 15% dane testowe. Symulacje wykonano w środowisku Matlab z pakietem Deep Learning, do nauki sieci użyto algorytmu Levenberga-Marquardta z funkcją straty będącą błędem średniokwadratowym MSE (*Mean Square Error*) estymacji współrzędnych punktu startowego. Inicjalizacja wag neuronów odbywała się z użyciem algorytmu Nguyena-Widrowa, który zakłada pewien stopień losowości, więc kolejne powtórzenia procesu uczenia mogą prowadzić do uzyskania sieci o różnej efektywności pracy, dlatego prezentowane dalej wyniki to najlepsze rezultaty z 30 powtórzeń procesu nauki sieci FNN.



Rys. 4. Regularne rozmieszczenie danych treningowych

Najlepsze rezultaty uzyskano stosując sieci, w których w warstwie lub warstwach ukrytych użyto funkcji aktywacji typu tangens hiperboliczny (tanh). Warstwa wejściowa z liniową funkcją aktywacji miała rozmiar 12, wynikający z liczby danych wejściowych, rozmiar warstwy wyjściowej, odpowiadał natomiast liczbie estymowanych współrzędnych czyli 2. Wyniki badań zbieżności estymacji położenia algorytmem genetycznym przy różnych rozmiarach warstw ukrytych przedstawiono w tab. 2, gdzie ε_{rms} i ε_{max} to odpowiednio błąd średniokwadratowy i maksymalny estymacji położenia przy użyciu samej sieci neuronowej, p_z to prawdopodobieństwo zbieżności algorytmu iteracyjnego, rozpoczynającego pracę od punktu o współrzędnych wskazanych przez sieć FNN, natomiast $k_{śr}$ i k_{max} to średnia i maksymalna liczba iteracji w algorytmie Gaussa-Newtona.

Analiza danych z tab. 2 dla sieci z jedną warstwą ukrytą pokazuje jak spada dokładność zgrubnego wyznaczania współrzędnych przy pomocy FNN przy zmniejszaniu rozmiaru warstwy z nieliniową funkcją aktywacji. Sieć zawierająca zaledwie 15 neuronów w warstwie ukrytej estymowała współrzędne z jakością wystarczającą do uzyskania zbieżności iteracyjnego algorytmu obliczania położenia we wszystkich 78961 punktach testowych. Rozbudowa sieci do dwóch warstw ukrytych pozwoliła na uzyskanie zbieżności w przypadku punktów startowych wskazywanych przez FFN z mniejszym rozmiarem warstw nieliniowych, ale minimalna łączna liczba neuronów w obu warstwach,

wymagana do prawidłowej pracy algorytmu iteracyjnego, wyniosła 11. Poniżej tej łącznej liczby neuronów żadna konfiguracja parametrów sieci FNN, trenowanej danymi generowanymi w sposób regularny, nie pozwoliła na pełną zbieżność algorytmu Gaussa-Newtona w całym obszarze badania usługi lokalizacyjnej. W takich przypadkach rozmieszczenie punktów, w których algorytm iteracyjny wykazywał zbieżność do nieprawidłowych współrzędnych, wskazywało na wyraźną regularność: obszary braku zbieżności występowały wyłącznie w bezpośrednim sąsiedztwie stacji bazowych. Należało zatem poprawić dokładność wskazywania pozycji startowej w obszarach blisko stacji bazowych, nawet kosztem pogorszenia efektywności FNN w innych obszarach.

Tab. 2. Zestawienie wybranych wyników analizy zbieżności przy stosowaniu sieci FNN z regularnym rozmieszczeniem danych treningowych

Rozmiar warstwy ukrytej	ε_{rms} *[m]*	ε_{max} *[m]*	p_z *[%]*	$k_{śr}$	k_{max}
Jedna warstwa ukryta (tanh)					
50	1,64	13,8	100	3,96	8
40	2,66	25,6	100	4,01	8
30	4,25	29,5	100	4,04	8
20	9,26	67,1	100	4,18	9
15	15,11	71,7	100	4,45	9
12	21,81	87,9	99,998	4,68	12
10	23,97	102,3	99,997	4,71	13
8	34,19	153,8	99,997	4,88	13
7	48,42	163,1	99,997	5,07	13
6	47,94	201,8	99,996	5,04	10
5	58,87	232,7	99,993	5,14	15
4	69,56	214,9	99,982	5,29	19
Dwie warstwy ukryte (tanh + tanh)					
6+5	16,84	82,7	100	4,47	10
6+6	8,73	46,9	100	4,17	12
10+4	10,46	55,6	100	4,28	10

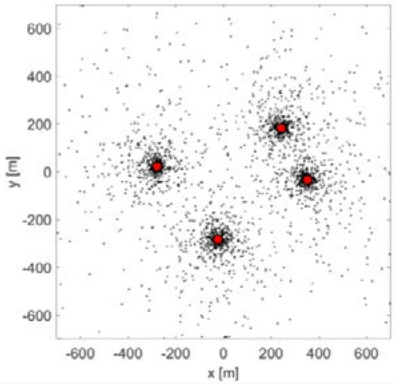
3.2. Nieregularne rozmieszczenie punktów do trenowania sieci neuronowej

W drugim etapie badań zaproponowano zmianę sposobu generowania danych treningowych: współrzędne kandydatów do zbioru treningowego są generowane losowo, po czym dany punkt jest dodawany do zbioru uczącego ze zmiennym prawdopodobieństwem p_d:

$$p_d = \max\left(1, \max\left(\frac{1}{d_n^2}\right)_{n=1\ldots N}\right), \qquad (5)$$

gdzie d_n jest odległością danego punktu od n-tej stacji bazowej. Przykład wygenerowanego zbioru z nierównomiernym rozmieszczeniem punktów do uczenia sieci przedstawiono na rys. 5.

Tak wygenerowany zbiór treningowy pozwolił na uzyskanie pełnej zbieżności iteracyjnego algorytmu estymacji położenia przy punktach startowych obliczanych przy użyciu sieci FNN z tylko jedną warstwą ukrytą, złożoną z zaledwie czterech neuronów. Sieć taka estymowała pozycję z błędem średniokwadratowym ε_{rms} wynoszącym aż 98,8m a błędem maksymalnym przekraczającym 340 metrów. Jednak w krytycznym obszarze w pobliżu wszystkich stacji bazowych żaden punkt startowy nie został wskazany po niewłaściwej stronie stacji bazowej, co pozwoliło na zbieżność algorytmu Gaussa-Newtona zawsze do właściwych współrzędnych wynikowych. Średnia liczba iteracji wyniosła 5,3 a maksymalna 8. Przy tak małej znalezionej strukturze sieci z tylko czterema neuronami o nieliniowej funkcji aktywacji, można uznać, że znaleziono minimalne rozwiązanie problemu wskazywania punktu startowego dla iteracyjnej estymacji położenia w systemie lokalizacji dwuwymiarowej przy użyciu sieci FNN.



Rys. 5. Przykład nieregularnego rozmieszczenia punktów do generowania danych treningowych dla sieci FNN

4. PODSUMOWANIE

Sztuczne sieci neuronowe mogą być użyte do realizacji zadań obliczeniowych bez ścisłego definiowania zależności matematycznych dla danego zadania. W niniejszym referacie sieć jednokierunkowa (FNN) została użyta do wskazywania punktu początkowego do pierwszej iteracji w iteracyjnym algorytmie estymacji położenia w systemie lokalizacji dwuwymiarowej. Znaleziona struktura sieci z tylko czterema neuronami z nieliniową funkcją aktywacji nie wymaga dużego nakładu obliczeniowego, a dodatkowo znacząca redukcja średniej i maksymalnej liczby iteracji, niezbędnych do estymacji położenia algorytmem Gaussa-Newtona w pełni kompensuje ten dodatkowy nakład przetwarzania realizowanego w sztucznej sieci neuronowej.

Następnym etapem prac może być modyfikacja sposobu uczenia sieci neuronowej w celu dalszej redukcji liczby iteracji wymaganych w procesie obliczania położenia terminala ruchomego.

LITERATURA

[1] Mensing C.,S. Plass. 2006. "Positioning Algorithms for Cellular Networks Using TDOA". *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Francja, pp. IV-IV, doi: 10.1109/ICASSP.2006.1661018.

[2] Foy W. 1976. "Position-Location Solutions by Taylor Series Estimation". *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-12, no. 2, str. 187-194

[3] Goodfellow I., Y. Bengio, A. Courville. 2016. "Deep Learning". MIT Press, online: http://www.deeplearningbook.org

Optymalizacja energetyczna szkieletu sieci z użyciem bypassów optycznych

Energy optimization of the network backbone employing optical bypasses

Wojciech Ząbek[1]; Piotr Jurkiewicz[2]; Robert Wójcik[3]; Jerzy Domżał[4]

[1] Akademia Górniczo-Hutnicza, Instytut Telekomunikacji, Kraków, email wojciech.zabek@agh.edu.pl
[2] Akademia Górniczo-Hutnicza, Instytut Telekomunikacji, Kraków, email piotr.jurkiewicz@agh.edu.pl
[3] Akademia Górniczo-Hutnicza, Instytut Telekomunikacji, Kraków, email robert.wojcik@agh.edu.pl
[4] Akademia Górniczo-Hutnicza, Instytut Telekomunikacji, Kraków, email jerzy.domzal@agh.edu.pl

DOI: 10.15199/59.2023.4.54

Streszczenie: Zaproponowane rozwiązanie przedstawia efektywne wykorzystanie optycznych zasobów sieciowych. Wykorzystuje w tym celu dobowe zmiany obciążenia sieci. Polega na częściowym wyłączaniu niewykorzystanych zasobów, z zachowaniem wymaganej przepływności sieci. Połączenia optyczne tworzone są tylko tam, gdzie są konieczne. Prezentowany mechanizm prowadzi do znacznego zmniejszenia zużycia energii elektrycznej, a jednocześnie nie powoduje degradacji przenoszonego ruchu.

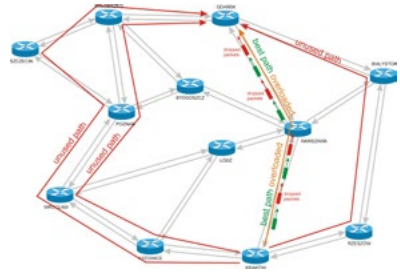
Abstract: The proposed solution shows efficient use of network optical resources. For this purpose, it uses daily network load changes. It consists in partial shutdown of unused resources, while maintaining the required network capacity. Connections are made only where necessary. The presented mechanism leads to a significant reduction in power consumption, and at the same time does not degrade the transferred traffic.

Słowa kluczowe: optyczne bypassy, przepływy, sieć IP, ruting, zużycie energii

Keywords: optical bypasses, flows, IP network, routing, energy consumption

1. WSTĘP

Sieci szkieletowe stoją przed wyzwaniem, jakim jest ciągły rozwój usług opartych o technologię IP. Rozwój jest przyczyną wzrostu ruchu IP. Duży przepływ danych w sieciach szkieletowych może być przyczyną powstawania zatorów. Jest to proces nagły i nieprzewidywalny. Zatory mogą powstać w dowolnym momencie i pomiędzy dowolnymi węzłami. Operator może zwiększyć przepustowości szkieletu sieci, ale proces ten jest czasochłonny, kosztowny i obciążający finansowo [5]. Stosowane obecnie mechanizmy w postaci protokołów OSPF, EIGRP i BGP mają wady. Wybór tylko jednej trasy nie jest najlepszym rozwiązaniem. Uniemożliwia pełne wykorzystanie zasobów sieci szkieletowej i sprzyja powstawaniu zatorów.



Rysunek 1: Problem protokołów rutingu, dot. wyboru najlepszej trasy.

Rysunek 1 ilustruje problem, pomiędzy węzłami Kraków – Gdańsk. Kiedy obciążenie między węzłami wzrasta, trasa ulega wysyceniu (pomarańczowa strzałka). Część pakietów może zostać porzuconych i/lub ulec opóźnieniu. Mechanizm okna TCP zmniejszy strumień danych i/lub ponowi utracone pakiety. Retransmisje zwiększą istniejący ruch – zmniejszą dostępną przepustowość. Pakiety UDP mogą zostać bezpowrotnie utracone. Zatory hamują także ruch sygnalizacyjny (proces trasowania).

Alternatywą dla rutingu klasycznego jest ruting wielościeżkowy. Umożliwia lepsze wykorzystanie zasobów i daje większą przepustowość między punktami końcowymi. W artykule proponujemy wykorzystanie rozszerzenia mechanizmu trasowania wielościeżkowego. Nowością jest wykorzystanie istniejących mechanizmów, tj. FAMTAR i AHB oraz połączenie ich funkcji celem zmniejszenia poboru mocy szkieletu sieci. W artykule zostanie omówione nowe podejście do mechanizmu zestawiania bezpośrednich połączeń w optyce. I tak rozdział 2 to opis problemu. Rozdział 3 prezentuje rozwiązanie, a środowisko badawcze i testy zawarte są w kolejnych. Ostatnia część, to wnioski wynikające z zastosowanego rozwiązania oraz z badań.

2. OPIS PROBLEMU

Rozważmy przypadek, w którym szkielet sieci jest mocno obciążony. Obciążenie jest zwykle nierównomierne. Pakiety z punktu A do punktu B pokonują zwykle najlepszą trasę w oparciu o system metryk i działający protokół rutingu dynamicznego. Metryki wyrażają domyślną przepustowość łączy, które tworzą całą ścieżkę od punktu A do punktu B. System ten umożliwia kierowanie pakietów najlepszą ścieżką, ale nie wykorzystuje w pełni dostępnych zasobów. Problemem jest niedoskonałość me-

chanizmu metryki. Wartości metryki reprezentują dostępną przepustowość pod względem statycznych kosztów. Nie dostarczają informacji o bieżącej zajętości poszczególnych łączy na trasie. Skutkuje to występowaniem blokad w najbardziej obciążonych węzłach/trasach. Protokoły rutingu dynamicznego mogą kierować pakiety kilkoma trasami jednocześnie. Mechanizm ten, oznaczony jako ECMP, wymaga tych samych kosztów tras równoległych, a także nie rozwiązuje problemu przeciążenia łączy. Ruting FAMTAR [4] (wyjaśniony w sekcji 2.1) umożliwia równoległe przesyłanie ruchu wieloma trasami oraz uwzględnia obciążenie łączy. FAMTAR dąży do najlepszej wydajności, ale nie bierze pod uwagę zużycia energii.

Artykuł [6] przedstawia rozwiązania umożliwiające lepsze wykorzystanie zasobów optycznych, ale wiele opcji redukcji zużycia energii nie wzięto w nim pod uwagę. Artykuł [8] porusza temat oszczędzania energii, ale nie uwzględnia trasowania w domenie optycznej. Problem śledzenia za pomocą optyki został omówiony w [3], gdzie zaproponowano algorytm oparty na funkcji kosztu, który waży stan zasilania elementów sieci w odniesieniu do wykorzystania długości fali. Wadą mechanizmu jest zwiększone prawdopodobieństwo zablokowania ruchu sieciowego. Rozwiązanie proponowane w tym artykule nie ma tej wady.

2.1 FAMTAR

Problem klasycznych protokołów rutingu dynamicznego przedstawiono na rysunku 1. Kiedy trasa między Krakowem, a Gdańskiem jest przeciążona (pomarańczowa strzałka), część pakietów jest porzucana. Oznacza to, że: alternatywne trasy nie są używane (czerwone strzałki), sieć jest nierównomiernie obciążona – zasoby nieefektywnie wykorzystane. Rozwiązaniem powyższego problemu jest mechanizm (metoda i algorytm) FAMTAR [7], który umożliwia kierowanie ruchu wieloma trasami w celu lepszego wykorzystania zasobów sieciowych.

Gdy sieć jest słabo obciążona, ruch odbywa się po najlepszych trasach. Ruting działa dokładnie tak, jak tradycyjny ruting dynamiczny. Jeżeli ruch na dowolnym łączu wzrośnie do zakładanego poziomu, nowe strumienie nie są kierowane na aktualnie wybraną trasę, aby nie powodować zatorów. Koszt łącza rośnie, a trasy stają się mniej atrakcyjne dla przyszłego ruchu. FAMTAR śledzi te zmiany i wybiera trasy o lepszych metrykach dla kolejnych, nowych przepływów. Stare przepływy nie są usuwane i kontynuują swoją pierwotnie wybraną ścieżkę do końca swojej transmisji. Kierowanie nowych przepływów trasami alternatywnymi ma pozytywny wpływ na ruch w szkielecie sieci. Zmniejsza zatory i umożliwia przesyłanie większej ilości danych w jednostce czasu w porównaniu z tradycyjnym rutingiem dynamicznym.

2.2 AHB

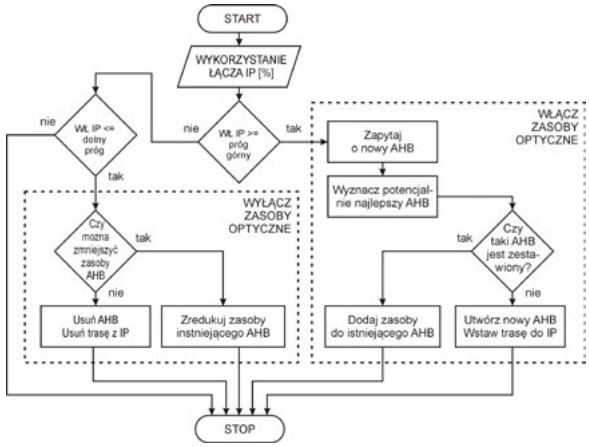
Mechanizm AHB został zaimplementowany i opisany w [2]. AHB jest algorytmem oraz alternatywą dla rozwiązania FAMTAR i częściowo powiela jego działanie, jednak operuje w domenie optycznej. Mechanizm umożliwia zestawienie bezpośredniego łącza optycznego z pominięciem warstwy elektrycznej węzłów pośredniczących. Używa do tego części zasobów optycznych zarezerwowanych dla AHB.

Kiedy na danym trakcie wzrośnie ruch do założonego poziomu, zestawiane jest dodatkowe połączenie. Nowe przepływy trafiają do tak zestawionego łącza celem ograniczenia ruchu w wysyconym trakcie. Kiedy ruch zmniejszy się do określonego poziomu, AHB jest likwidowane, a przepływy przesyłane domyślnym traktem.

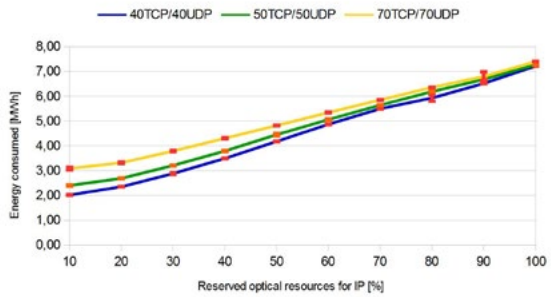
3. SAMODOSTOSOWUJĄCA SIĘ SIEĆ

Korzystne jest zarezerwowanie jak największej ilości zasobów sieci optycznej dla AHB, przy zachowaniu minimalnych zasobów dla IP [1]. Gdy obciążenie sieci szkieletowej jest niskie, część zasobów transpondera optycznego jest wyłączona, co oznacza redukcję zużycia energii [9] oraz więcej wolnych zasobów dla przyszłych przepływów. Wolne zasoby umożliwiają dynamiczne konfigurowanie logicznej topologii szkieletu sieci, z uwzględnieniem przyszłego ruchu.

Algorytm przedstawiony na rysunku 2, uruchamiany jest cyklicznie co 60 sekund. Po wywołaniu iteruje po wszystkich łączach IP i sprawdza ich zajętość. Jeśli łącze wysycone jest ruchem powyżej pewnego progu, następuje zestawienie nowego AHB lub już istniejący zasób AHB zostaje rozszerzony (w %). W przypadku nowego AHB informacja o łączu trafia do warstwy IP (FAMTAR). Jeśli istniejący AHB osiągnie niższy poziom nasycenia łącza, przepustowość AHB jest zmniejszana poprzez odejmowanie zasobów. AHB jest likwidowany, gdy osiągnie wartość minimalną i przepływ spadnie poniżej założonego progu. AHB może być zestawiany dowolnie pomiędzy sąsiednimi węzłami oraz z pominięciem węzłów tranzytowych.



Rysunek 2: Algorytm ustanawiania i redukcji AHB. WŁ - wykorzystanie łącza IP.



Rysunek 3: Zależność pomiędzy liczbą zasobów zarezerwowanych dla IP, a całkowitym zużyciem energii.

4. ŚRODOWISKO BADAWCZE

Badania przeprowadzono w środowisku MiniNet, wraz z zewnętrznym kontrolerem SDN. Symulowano działanie sieci w topologii Polskiej sieci szkieletowej.

W testowanej topologii zasoby każdego łącza optycznego przydzielane są procentowo dla ruchu IP. Pozostałe, nieprzydzielone zasoby nie są dostępne dla IP, nie przesyłają ruchu i są dostępne dla AHB. Nie przydzielone zasoby oznaczają wyłączenie części transponderów optycznych, co daje zmniejszeniem poboru mocy przez szkielet sieci.

Ruch w symulatorze MiniNet generowany jest losowo przez generatory TCP i UDP, w zakresie 5-200 Mbit/s. Generator UDP zapewnia dobową obwiednię, gdzie minimum przypada na pory nocne, a maksimum w dzień. Generator TCP zapewnia różne rodzaje natężenia szczytowego ruchu TCP. Założenia wynikają ze względu na ograniczenia symulatora.

Tabela 1: Parametry wejściowe symulacji.

Parametr	*Wartość*	*Jednostka*
FAMTAR	zajętość łącza	
aktywacja	85	[%]
deaktywacja	75	[%]
AHB	zajętość łącza	
aktywacja	98	[%]
deaktywacja	50	[%]
Czas symulacji	5	[godziny]
Liczba węzłów	12	szt.
Liczba tras	18	szt.
Maks. przep. traktu	100	[Mbit/s]
Maks. gen. TCP	40/50/70	[Mbit/s]
Maks. gen. UDP	40/50/70	[Mbit/s]
Zasoby dla IP	10-100	[%]

Tabela 1 przedstawia kluczowe parametry symulacji. Dotyczą one ustawienia warunków pracy symulatora, tj. ilość generowanego ruchu TCP/UDP, wstępnej alokacji zasobów do warstwy IP itp. Podobnie jak AHB, FAMTAR aktywowany i dekatywowany jest przy określonej, procentowej zajętości traktu/łącza.

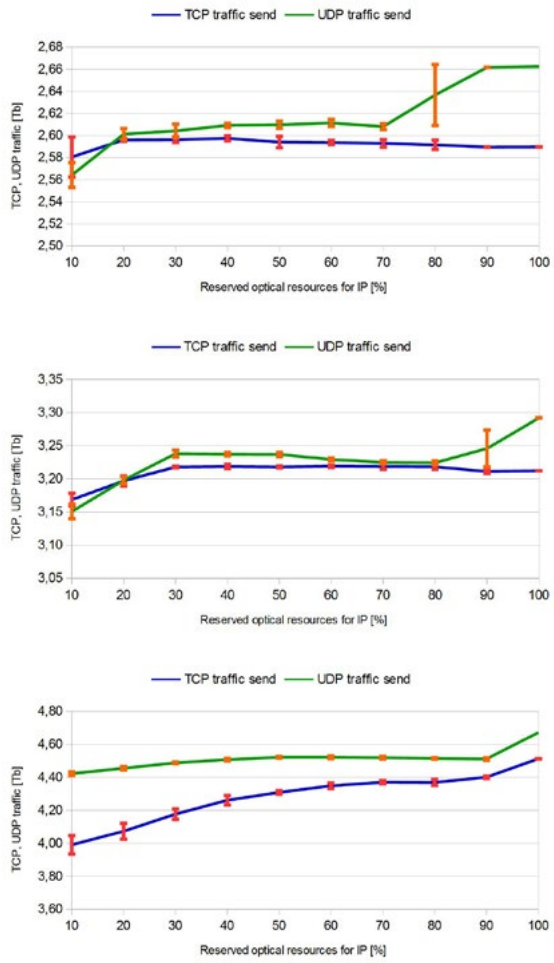
5. TESTY

Przeprowadzono kilkaset 5-godzinnych symulacji. W każdej symulacji pierwsze dwie godziny nie były brane pod uwagę do dalszej analizy (czas rozbiegowy). Każdą symulację powtarzano siedem razy. Otrzymane wartości obliczono dla parametrów ruchu sieciowego na podstawie rozkładu t-Studenta (poziom istotności równy 0,05). Przeprowadzone badania wykazały, że:

- przy najmniejszej ilości zasobów zarezerwowanych dla IP, szkielet sieci charakteryzował się znacznymi oszczędnościami w zużyciu energii elektrycznej,
- sumaryczny transfer UDP został zwiększony.

W symulacjach ustalono maksymalną ilość generowanych danych (generatory TCP/UDP). Przyjęto je na poziomie 40, 50 i 70 Mbit/s każda (tabela 1). Scenariusz z 40/40 Mbit/s reprezentuje łącza niewykorzystane (każde łącze ma przepustowość 100 Mbit/s). Scenariusz z 50/50 Mbit/s pokazuje zachowanie nasyconej sieci. Scenariusz 70/70 Mbit/s przedstawia sieć mocno przeciążoną. Parametry symulacji można dowolnie skalować do przepustowości łączy między węzłami, np. 1 Gbit/s, 10 Gbit/s i 100 Gbit/s. Zależność zużytej energii elektrycznej od zasobów zarezerwowanych dla IP pokazuje rysunek 3.

Na rysunku 3 przedstawiono trzy linie, które odpowiadają wartościom szczytowym generatorów. Z wykresu wynika, że dla wartości 10, 20 i 30% wykorzystanych zasobów dla IP zużycie energii elektrycznej jest znacznie mniejszesze i sięga połowie wartości w porównaniu do 100% wykorzystanych zasobów dla IP (100 Mbit/s). Rysunek 3 pokazuje również, że zużycie energii elektrycznej jest maksymalne, gdy zasoby zarezerwowane dla IP są na poziomie 90-100% (około 7 MWh). Gdy zarezerwowane zasoby dla IP są na poziomie 20-30%, zużywana energia elektryczna jest prawie o połowę mniejsza. Ilość przesyłanych danych TCP i UDP praktycznie się nie zmienia i waha się w granicach błędu statystycznego.

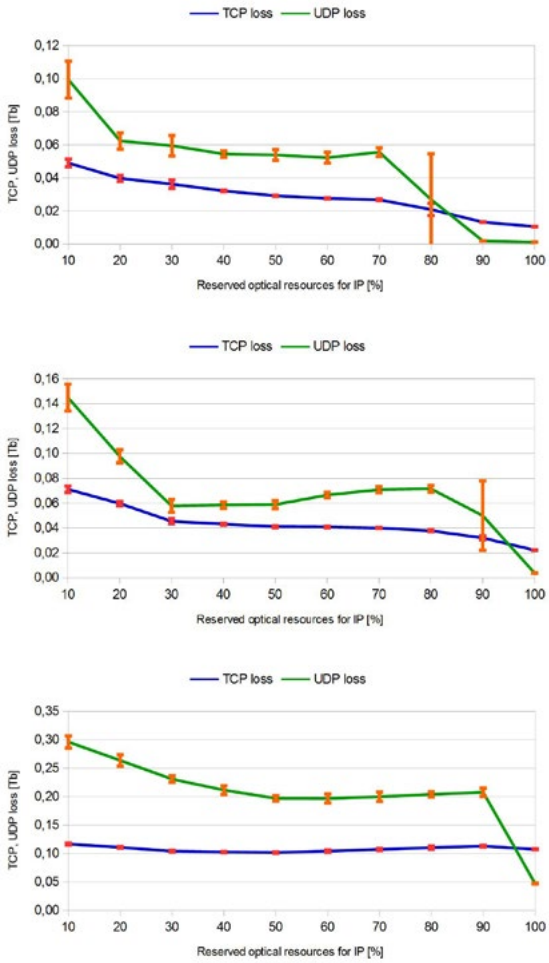


Rysunek 4: Zależność pomiędzy liczbą zasobów zarezerwowanych dla IP, a całkowitą liczbą przesłanych danych TCP/UDP (generatory odpowiednio od góry 40/40, 50/50 oraz 70/70).

Rysunek 4 przedstawia transfery uzyskane za pomocą generatorów TCP/UDP osiągające maksymalnie odpowiednio 40/50/70 Mb/s dla TCP i 40/50/70 Mb/s dla UDP. Przy wartościach zarezerwowanych zasobów dla IP na poziomie 20-70% suma przesyłanych danych jest praktycznie niezmienna. Wartość ta nieznacznie wzrasta dla ruchu UDP (zielona linia), gdy zasoby zarezerwowane dla IP przekraczają 70%. Rysunek pokazuje anomalię przepływu UDP dla progu odpowiednio 80 i 90% (generatory odpowiednio 40/40 oraz 50/50). Wynika to z nieliniowości

procesu włączania i wyłączania FAMTAR z maksymalną możliwą szybkością. W efekcie występują większe wahania uzyskanych wyników w stosunku do przyjętego przedziału ufności. Fluktuacje nie występują dla TCP (ciemnoniebieska linia), ze względu na mechanizm okna oraz dla generatorów 70/70, ze względu na wyższe uzyskane przepływności. Przy sieci przeciążonej (generatory 70/70 Mb/s), wspomniana wyżej nieliniowość nie występuje.

Rysunek 5 pokazuje straty pakietów TCP i UDP w funkcji rezerwacji zasobów dla IP.



Rysunek 5: Zależność pomiędzy liczbą zasobów zarezerwowanych dla IP, a całkowitą utratą przesyłanych pakietów TCP/UDP (generatory odpowiednio od góry 40/40, 50/50 oraz 70/70).

Rysunek 5 przedstawia wykres symulacji dla generatorów TCP i UDP do wartości odpowiednio 40/50/70 Mb/s. Utracone pakiety kształtują się na podobnym poziomie, tj. w przedziale 20-70% zasobów zarezerwowanych dla IP. Powyżej odpowiednio 70/80/90% straty maleją i dążą do zera, gdy zasoby dla IP wynoszą 90-100%. W obrębie odpowiednio 80/90%, przy generatorach odpowiednio 40/50 Mb/s, zasobów zarezerwowanych dla IP występuje wcześniej opisana anomalia. Spadek utraty danych jest większy w przypadku ruchu UDP.

6. WNIOSKI

Badany szkielet sieci nie zawsze jest jednakowo obciążony. Obciążenie poszczególnych obszarów zmienia się w cyklu dobowym. Dynamiczne zarządzanie dostępnymi zasobami sieci optycznej jest rozsądną alternatywą, gdyż część zasobów może zostać wyłączona (mały ruch). Rozwiązanie to wprowadza oszczędności energii elektrycznej zasilającej urządzenia aktywne. Kiedy ruch wzrasta, szkielet sieci może użyć wolne zasoby do zestawiania bezpośrednich połączeń optycznych AHB między wybranymi węzłami. Zaproponowany algorytm jest prosty, elastyczny, mniej podatny na awarie i oszczędzający zasoby. Ponadto AHB tworzy bezpośrednie połączenia optyczne, z pominięciem IP. Różnica polega na podejściu – szkielet sieci konfiguruje wymagane połączenia w razie potrzeby i nie ogranicza już istniejących.

Zastosowana koncepcja redukcji użycia zasobów (AHB) jest dobrym kierunkiem. Sieć szkieletowa z możliwością samokonfiguracji, jest lepszą alternatywą dla redukcji kosztów eksploatacyjnych, co może przełożyć się na ceny usług oraz zwiększone zyski.

LITERATURA

[1] Biernacka Edyta, Boryło Piotr, Wójcik Robert, Domżał Jerzy. 2019. „Elastic optical bypasses for traffic bursts". Computer Communications, 146 : 360–393.

[2] Domżał Jerzy, Duliński Zbigniew, Rząsa Jacek, Gawłowicz Piotr, Biernacka Edyta, Wójcik Robert. 2017. „Automatic hidden bypasses in software-defined networks". Journal of Network and Systems Management, 25 (3) : 457–480.

[3] Idzikowski Filip, Orlowski Sebastian, Raack Christian, Woesner Hagen, Wolisz Adam. 2012. „Power savings versus network performance in dynamically provisioned WDM networks". 2010 14th Conference on Optical Network Design and Modeling (ONDM), 50 : 48–50.

[4] Jurkiewicz Piotr, Wójcik Robert, Domżał Jerzy, Kamisiński Andrzej. 2020. „Testing implementation of FAMTAR: Adaptive multipath routing". Computer Communications, 149 : 300–311.

[5] Marcus J. S. 2014. „The Economic Impact of Internet Traffic Growth on Network Operators". wIK-Consult Google Inc.

[6] Tanaka Takafumi, Tetsuro Inui, Kadohata Akihiro, Imajuku Wataru, Hirano Akira. 2016. „Multiperiod IP-over-elastic network reconfiguration with adaptive bandwidth resizing and modulation". Journal of Optical Communications and Networking, 8.

[7] Wójcik Robert, Domżał Jerzy, Duliński Zbigniew. 2014 „Flow-Aware Multi-Topology Adaptive Routing". IEEE Communications Letters, 18 (9) : 1539–1542.

[8] Xia Ming, Massimo Tornatore, Zhang Yi, Chowdhury Pulak, Martel Charles U., Mukherjee Biswanath. 2011. „Green Provisioning for Optical WDM Networks". IEEE Journal of Selected Topics in Quantum Electronics, 17.

[9] Zhang Jiawei, Zhao Yongli, Yu Xiaosong, Zhang Jie, Song Mei, Ji Yuefeng, Mukherjee Biswanath. 2015. „Energy-Efficient Traffic Grooming in Sliceable-Transponder-Equipped IP-Over-Elastic Optical Networks [Invited]". 2009 International Conference on Intelligent Networking and Collaborative Systems, 7 : A142–A152.

ODPORNOŚĆ SIECI xPON NA DZIAŁNIE NIEPOSŁUSZNEJ JEDNOSTKI ONU

RESISTANCE OF xPON NETWORKS TO THE OPERATION OF ROGUE ONU

Sławomir Kula[1]; Weronika Kowalska[2]

[1] Instytut Telekomunikacji Politechnika Warszawska, slawomir.kula@pw.edu.pl
[2] Wydział Elektroniki i Technik Informacyjnych Politechniki Warszawskiej, weronika.kowalska8@pw.edu.pl

DOI: 10.15199/59.2023.4.55

Streszczenie: **Referat jest poświęcony problemowi zakłócania lub uniemożliwiania transmisji w górę w sieciach dostępowych xPON w wyniku niewłaściwego działania jednostki ONU, tak zwanej jednostki rogue ONU wchodzącej w skład sieci xPON. W tym celu przeprowadzono badania laboratoryjne sieci GPON, zastępując jedną z jednostek ONU przestrajanym nadajnikiem laserowym. Uzyskane wnioski rozciągnięto na systemy XG-PON i NG-PON.**

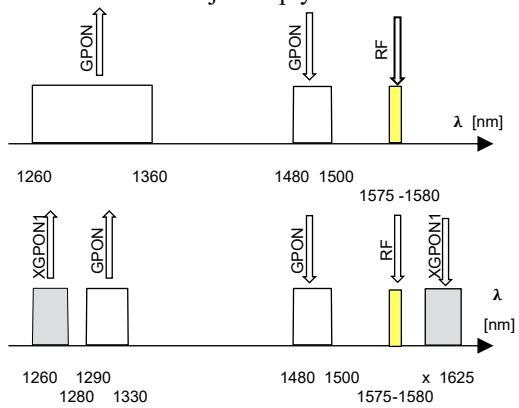
Abstract: **The paper is devoted to the problem of interfering or preventing upstream transmission in xPON access networks as a result improper operation of the ONU, the so-called rogue ONU included in the xPON network. For this purpose laboratory tests of GPON network were carried out, replacing one of the ONUs with a tunable laser transmitter. The obtained conclusions were extended to XG-PON and NG-PON networks.**

Słowa kluczowe: **GPON, XG-PON, NG-PON sieci dostępowe, rogue ONU.**

Keywords: **GPON, XG-PON, NG-PON access networks, rogue ONU.**

1. WSTĘP

Pasywne optyczne sieci dostępowe xPON (x *Passive Optical Network*) podlegają bardzo dynamicznemu rozwojowi. Dzieje się tak z kilku powodów, ale najważniejszy z nich to olbrzymi wzrost zapotrzebowania na szybkość transmisji w warstwie dostępowej spowodowany nie tylko przez klientów indywidualnych, ale także przez możliwość wykorzystywania sieci xPON do transmisji sygnałów ze stacji bazowych sieci 5G. Po sieciach standardu BPON [1] pojawiły się sieci standardu GPON [2], [3] sieci XG-PON [4] i sieci następnej generacji NG-PON2 [5]. W większości krajów, także i w Polsce najpowszechniej są stosowane sieci standardu GPON, a sieci najnowszych standardów dopiero wchodzą do eksploatacji. Warto jednak pamiętać, że migracja w kierunku nowych rozwiązań może powodować problemy, jeżeli w optycznej sieci dystrybucyjnej ODN (*Optical Distribution Network*) mają pracować równocześnie jednostki ONU (*Optical Network Unit*), albo optyczne jednostki zakończenia sieci ONT (*Optical Network Termination*) różnych standardów, a taki scenariusz migracji bierze pod uwagę wielu operatorów na świecie. Aby to było możliwe odpowiednio przydzielono widmo poszczególnym standardom, tak by widma nie pokrywały się. Wyjątkiem są standardy GPON i XG-PON, a dokładniej jednostki ONU/ONT starszej generacji sieci GPON, które do transmisji w górę wykorzystują lasery z rezonatorami Fabry-Perot [2] – lasery MLM (*MultiLongitudinal Mode*). Pierwotnie w sieciach tych na potrzeby transmisji w górę przeznaczono pasmo od 1260 nm do 1360 nm. Wprowadzenie standardu XG-PON1 spowodowało konieczność podzielenia tego pasma na pasmo od 1260 nm do 1280 nm używane przez systemy XG-PON1 i zawężenie pasma dla systemów GPON do zakresu od 1290 nm 1330 nm [3] (Rys.1). Praca w tym zakresie przez systemy GPON jest możliwa tylko w przypadku stosowania laserów ze sprzęgaczem zwrotnym DFB (*Distributed FeedBack*) – SLM (*SingleLongitudinal Mode*). Dalsza redukcja do przedziału 1300÷1320 nm jest możliwa tylko wtedy, gdy w jednostkach ONU/ONT GPON stosuje się wyselekcjonowane laserowe nadajniki optyczne SLM.



Rys. 1. Zakresy długości fal wykorzystywane przez systemy GPON i XG-PON1 (rysunek górny dotyczy starszej wersji systemów GPON)

Przyjęcie takiego rozwiązania uniemożliwia równoczesne użytkowanie w sieci ODN jednostek GPON starszego typu i jednostek XG-PON. Optyczne jednostki sieciowe ONU/ONT, standardu GPON muszą w takim przypadku mieć nadajniki laserowe pracujące w węższym paśmie. Dla czytelności w dalszej części referatu odnosząc się do jednostki abonenckiej używany będzie tylko skrót ONU, ponieważ różnica między jednostkami ONU i ONT nie ma żadnego znaczenia z punktu widzenia niniejszego referatu. W starszych, ale wciąż eksploatowanych sieciach GPON używanie laserów z rezonatorem Fabry-Perot'a grozi niebezpieczeństwem zablokowania transmisji w górę w sytuacji, gdy jedna z jednostek ONU (tzw. nieposłuszna – *rogue*) z jakichś powodów zacznie nadawać poza przydzielonymi jej przez jednostkę OLT (*Optical*

Line Termination) slotami czasowymi. W rezultacie dojdzie do nakładania się impulsów światła nadawanych przez poprawnie działające jednostki ONU z impulsami nadawanymi przez nieposłuszną jednostkę ONU i zablokowania działania sieci. W referacie [6] omówiono szczegółowo problem nieposłusznych jednostek ONU i zaproponowano metodę jego rozwiązania w przypadku, gdy jednostki ONU używają laserów MLM. Lasery tego typu mogą być używane tylko przez jednostki ONU GPON (Tab.1).

Tab. 1. Typy laserów stosowane w sieciach xPON

	GPON	*XG-PON*	*NG-PON*
w dół	SLM	SLM	SLM
w górę	MLM/SLM	SLM	SLM

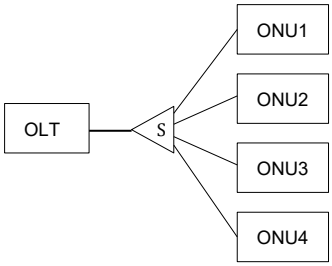
Otwartym pozostaje pytanie: czy gdy są używane nadajniki laserowe emitujące zdecydowanie bardziej wąskopasmowe światło zablokowanie transmisji w górę przez nieposłuszną jednostkę ONU może wystąpić? Praktyczne sprawdzenie tego jest głównym, ale nie jedynym celem referatu. Na tej podstawie oraz na podstawie analizy standardów rozszerzono wyciągnięte z badania sieci GPON wnioski na systemy XG-PON i NG-PON.

2. STANOWISKO TESTOWE

Testy sieci GPON przeprowadzono w Laboratorium Systemów Dostępowych i Teletransmisyjnych Instytutu Telekomunikacji Politechniki Warszawskiej.

2.1. Struktura testowanej sieci GPON

Testowana sieć GPON miała strukturę gwiazdy (Rys.2). Składała się z: światłowodowej sieć dystrybucyjnej ODN, której elementem był splitter 1x4, cztery jednostki ONU i jednostka OLT.



Rys.2. Podstawowa struktura testowanej sieci

2.2. Urządzenia testowanej sieci GPON

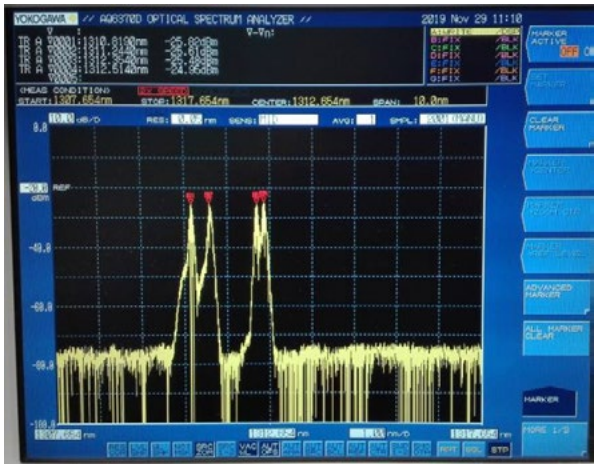
W skład sieci GPON wchodziła jednostka OLT Zhone MXX-19x1U [7], cztery jednostki ONU Zhone ZNID-GPON-2424a [8] oraz splitter optyczny 1x4 GPM-104-PS-30-10-44. Światłowody jednomodowe sieci ODN miały długości nieprzekraczające kilkunastu metrów. Nominalne długości fal używanych w teście jednostek ONU, a także dodatkowego sterowanego nadajnika laserowych zostały zmierzone za pomocą analizatora widma optycznego YOKOGAWA AQ63700, którego dokładność w zakresie drugiego i trzeciego okna optycznego jest lepsza niż ± 0,1 nm.

Według specyfikacji producenta nominalna długość fali nadajników optycznych jednostki ONU zawiera się w przedziale 1290-1330 nm i są nimi lasery SLM. Moc sygnału nadawanego w górę może być zmieniana w zakresie od 0,5 dBm do 6 dBm. W Tabeli 2 podano zmierzone długości fal nadawanych przez jednostki ONU.

Tab. 2. Długości fal emitowanych przez jednostki ONU

Nr jednostki ONU	*Długość fali [nm]*
1	1310,819
2	1311,244
3	1312,354
4	1312,514

Na Rys.3 pokazano zrzut z ekranu analizatora widma, obrazujący widma sygnałów nadawanych przez jednostki ONU.

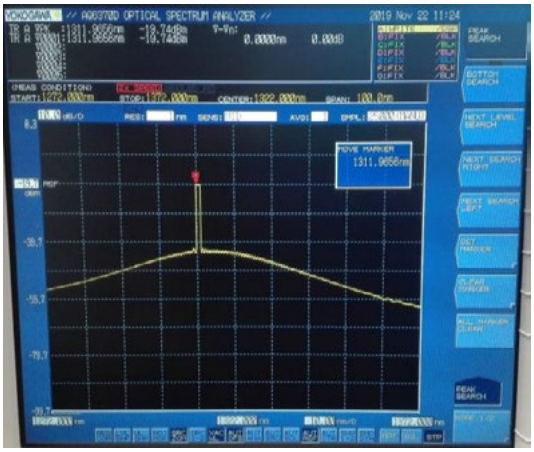


Rys. 3. Widma sygnałów nadawanych przez jednostki ONU (uwaga: na ekranie analizatora oznaczane jako jednostki ONT) [9]

Ustawione moce sygnałów nadawanych przez jednostki ONU były teoretycznie takie same, choć w rzeczywistości mogły się nieznacznie różnić. To, a także różnica w tłumieniu sygnałów na stykach rozłączalnych spowodowała, że zmierzone moce sygnałów nie były jednakowe. Różnica nie przekraczała 1 dBm.

Według rekomendacji ITU-T w przypadku używania do transmisji w górę widma z zakresu 1290÷1330 nm konieczne jest stosowanie laserów SLM, których widmo sygnałów nadawanych mierzone 20 dB poniżej maksymalnej wartości ma szerokość nie przekraczającą 1 nm. Widmo światła emitowanego przez lasery SLM jest zwykle znacznie węższe, ale należy pamiętać, że lasery używane w jednostkach ONU GPON, ze względu na minimalizację ich ceny nie są wysokiej jakości, w szczególności nie mają stabilizacji temperatury. Dotyczy to także jednostek ONU używanych w testowanej sieci.

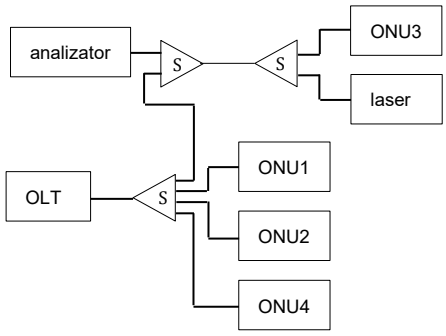
Jednostka nieposłuszna ONU była symulowana za pomocą lasera PRO800 Laser Diode Current and Temperature Control. Zmieniając temperaturę i natężenie prądu można odpowiednio zmienić moc i długość fali nadawanego sygnału optycznego. Na Rys.4 pokazano przykładowe widmo sygnału nadawanego przez laser PRO800.



Rys. 4. Widmo sygnału optycznego lasera PRO800 (I=4,7 mA, T=0,3 ° C) [9]

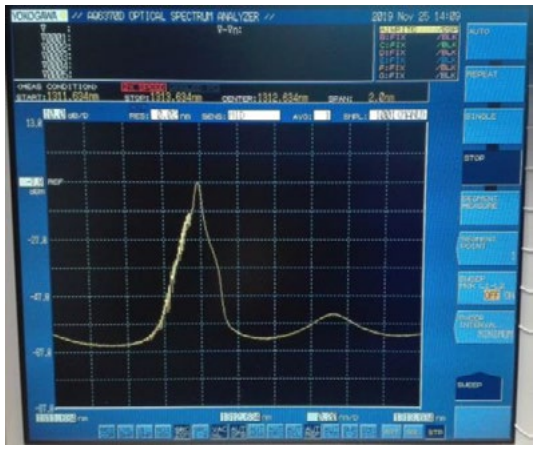
3. PRZEBIEG I WYNIKI TESTÓW

W celu zbadania zachowania się sieci GPON w przypadku, gdy jedna z jednostek ONU zmienia swój status na nieposłuszną uzupełniono sieć o dwa splittery 1x2 (Rys.5). Jednostkę nieposłuszną symulowano za pomocą lasera PRO 800. Dzięki takiej konfiguracji sieci możliwe było równoczesne analizowanie widma sygnałów nadawanych przez laser i jednostkę ONU3. Analiza sygnalizacji alarmowej jednostki OLT pozwalała na obserwowanie wpływu lasera na transmisję w górę nie tylko z jednostki ONU3, ale także pozostałych trzech jednostek. Na Rys.5 nie zaznaczono dodatkowych tłumików, których umieszczenie było konieczne by moce sygnałów docierających do jednostki OLT mieściły się w dopuszczalnym zakresie.



Rys. 5. Konfiguracja sieci w trakcie testów

Zmieniając temperaturę lasera zmieniano długość fali i ustawiono ją na taką samą z jaką pracowała jednostka ONU3, czyli 1312,424 nm (Rys.6). Uwaga, ponieważ w jednostkach ONU testowanej sieci nie ma stabilizacji temperatury długość fali nadawanej przez tę jednostkę różni się od podanej w Tab.1. Moc sygnału nadawanego przez laser była stopniowo zwiększana, aż do momentu, gdy jednostka OLT zaczęła sygnalizować degradację jakości SD (*Signal Degraded*) sygnału odbieranego od jednostki ONU3, co oznaczało przekroczenie bitowej stopy błędów, która była ustawiona na 10^{-8} [10]. Laser nie wpływał na jakość sygnałów i pracę pozostałych jednostek ONU mimo, że pracował on w trybie ciągłym, a więc nadawał także w szczelinach czasowych dla nich przeznaczonych. Oznacza to, że jednostka OLT odbiera sygnały optyczne selektywnie, za każdym razem dostrajając odbiornik do długości fali, na której nadaje dana jednostka ONU. Dalsze zwiększanie mocy lasera powodowało pojawianie się kolejnych alarmów i w ostateczności wskazania defektu po stronie zdalnej RDI (*Remote Defect Indication*) dotyczącego jednostki ONU3. Alarm RDI automatycznie blokuje transmisję od i do tej jednostki.



Rys. 6. Widmo sygnału optycznego jednostki ONU3 i lasera PRO800 (I=10 mA, T=0,05 ° C) [9]

Co ważne, stało się tak dopiero, gdy moc sygnału z lasera przekroczyła moc sygnału emitowanego przez jednostkę ONU3 o 9,5 dB. Wyłączenie i ponowne włączenie zasilania jednostki ONU3, przy niezmienionej mocy nadajnika lasera przywróciło poprawną, bez alarmów pracę jednostki ONU3. Aby wyjaśnić oba te zaskakujące zachowania sieci należy odwołać się do procedury odbioru danych (paczki danych) z jednostki ONU w dedykowanym dla niej slocie czasowym. Wartościom logicznym bitów odpowiadają różne poziomy mocy odbieranego sygnału. Różnica średnich poziomów mocy między bitem „1" a bitem „0" powinna wynosić co najmniej 20 dB i jest określana przez parametr EX (*Extinction Ratio*) [10]. Sygnał z nadającego stale lasera dodaje się w jednakowym stopniu do nadawanych „1" jak i „0", przez co ich rozróżnienie nie stanowi problemu. Jeżeli jednak łączny poziom mocy odbieranego sygnału przekroczy wartości uzgodnione w fazie uruchamiania jednostki ONU, to spowoduje pojawienie się alarmu blokującego transmisję. Uzgadnianie i ewentualna zmiana poziomu mocy sygnału nadawanego przez jednostkę ONU w trakcie normalnej pracy nie jest możliwa. Po wyłączeniu i ponownym włączeniu zasilania jednostki ONU3 następuje procedura ponownego ustalania mocy i dopasowania poziomów do nowych warunków. Moc sygnału nadawanego przez jednostkę ONU może być zmieniona o ± 3 dBm [10].

Na zakończenie tego punktu należy podkreślić, że w praktyce, aby problem nieposłusznej jednostki wystąpił, muszą być spełnione równocześnie trzy warunki: widmo nadajnika nieposłusznej jednostki musi pokrywać się z widmem wykorzystywanym przez inną jednostkę ONU, jednostka nieposłuszna musi nadawać w szczelinach czasowych jej nieprzydzielonych i wreszcie moc sygnału nadawanego przez nieposłuszną jednostkę musi być odpowiednio duża. Spełnienie tych warunków jest możliwe w

wyniku awarii jednostki ONU, lub awarii i zmiany temperatury nadajnika jednostki, ale jest mało prawdopodobne.

4. NIEPOSŁUSZNE JEDNOSTKI W SIECIACH XG-PON i NG-PON2

Analiza standardów ITU-T dotyczących systemów XG-PON i NG-PON2 pozwala wykorzystać wyniki przeprowadzonych testów do oceny niebezpieczeństwa spowodowanego nieposłuszną jednostką ONU w sieciach tych standardów o strukturze gwiazdy.

4.1. XG-PON

W sieciach XG-PON o strukturze gwiazdy niebezpieczeństwo zablokowania transmisji przez nieposłuszną jednostkę jest możliwe, i tak samo prawdopodobne jak w sieciach GPON, bowiem w obu przypadkach nadajniki laserowe w jednostkach ONU są typu SLM bez stabilizacji temperatury. Awaria jednostki ONU i równoczesne przesunięcia widma emitowanego sygnału tak, że nałoży się ono na widmo używane do transmisji w górę przez inną jednostkę może jednak wystąpić.

4.2 NG-PON2

W sieciach NG-PON2 również stosuje się nadajniki laserowe SLM, ale ze względu na wprowadzenie multipleksacji TWDM **(***Time and Wavelenght Division Multiplexing channels*) i transmisji PtP WDM (*Point to Point WDM*) problem jednostek nieposłusznych może w ogóle nie występować w przypadku transmisji PtP WDM, a w przypadku TWDM jego wystąpienie może być bardzo mało prawdopodobne. W transmisji PtP jednostka ONU i jednostka OLT nadają na różnych długościach fal dedykowanych wyłącznie dla danego klienta (danej jednostki ONU). Z kolei zastosowanie rozwiązania TWDM oznacza, że jedna długość fali dla kierunku w dół a inna dla kierunku w górę może być używana przez kilka jednostek ONU. W takiej sytuacji są przydzielane sloty czasowe tak jak we wcześniejszych standardach. Jest jednak pewna istotna różnica. Długości fal są ściśle określone (z dokładnością do dwóch miejsc po przecinku, gdy długość fali jest podawana w nm). Aby wykluczyć niebezpieczeństwo zmiany długości fali, na której nadaje jednostka ONU NG-PON2 powinny być stosowane nadajniki laserowe ze stabilizacją temperatury,

5. WNIOSKI

Wykonane badania potwierdzają, że nadawanie przez jednostkę ONU (nieposłuszną jednostkę ONU) w sieci dostępowej GPON niekoniecznie będzie skutkowało blokadą transmisji przez pozostałe jednostki ONU. W przypadku sieci GPON, w których jednostki ONU używają nadajników SLM nie będzie zatem dochodziło do blokady wszystkich jednostek, która może wystąpić w sieciach GPON korzystających z pasma od 1260 nm do 1360 nm i nadajników optycznych MLM w jednostkach ONU. W sieciach GPON może jednak też dojść do sytuacji, w której nieposłuszna jednostka zablokuje możliwość korzystania z innej jednostki ONU. Aby tak się stało sieć musi mieć strukturę gwiazdy albo podwójnej gwiazdy, nieposłuszna jednostka ONU, w wyniku niestabilnej pracy musiałaby nadawać na długości fali innej jednostki, a dodatkowo nieposłuszna jednostka musi nadawać z odpowiednio dużą mocą. Jeżeli łączna moc sygnałów z obu jednostek nie powoduje przesterowania, to wyłączenie i ponowne włączenie zasilania zakłócanej jednostki ONU może przywrócić jej poprawną pracę. Tak samo będzie w przypadku sieci XG-PON. Negatywne skutki, gdy jedna z jednostek ONU stanie się nieposłuszna w sieci NG-PON2 o strukturze gwiazdy albo podwójnej gwiazdy są jeszcze mniej prawdopodobne niż w sieciach GPON i XG-PON. Jest tak, ponieważ lasery SLM używane w jednostkach ONU sieci NG-PON2 mają stabilizatory temperatury i ewentualnie tylko uszkodzenie zespołu stabilizującego temperaturę lasera może negatywnie wpływać na pracę innych jednostek.

LITERATURA

[1] ITU-T Rec. G.983.1 (1998). “Broadband optical access systems based on Passive Optical Networks (PON)”.
[2] ITU-T Rec. G.984.1 (2003). “Gigabit-capable Passive Optical Networks (GPON): Physical Media Dependent (PMD) layer specification".
[3] ITU-T Rec. G.984.5 (05/2014). “Gigabit-capable passive optical networks (GPON): Enhancement band”.
[4] ITU-T Rec. G.987.1 (2010). “10-Gigabit-capable passive optical networks (XG-PON): General requirements”.
[5] ITU-T Rec. G.989.1 (2013). “40-Gigabit-capable passive optical networks (NG-PON2): General requirements”.
[6] Kula Sławomir. 2019. „Metoda wykrywania Rogue-ONU - nieposłusznej jednostki ONU w sieci GPON”. KSTiT.
[7] „Baza OLT Zhone MXK-194, 4xGPON, 8 FE/GE (SFP), 1U19” - specyfikacja techniczna. https://www.kabeltechnika.pl/produkt/2427/baza-olt-zhone-mxk194-4xgpon-8-fege-sfp-1u.html
[8] „Terminal abonencki ONT Zhone 2426, 4xGE, 2xPOTS, WiFi” - specyfikacja techniczna. https://www.kabeltechnika.pl/produkt/2433/terminal-abonencki-ont-zhone-2426-4xge-2xpots.html
[9] Grzegory Weronika. 2020. „Badanie odporności sieci GPON na występowanie w niej jednostek „rogue-ONU”. Praca inżynierska. WEITI PW.
[10] ITU-T Rec. G.984.3 (2008). “Gigabit-capable Passive Optical Networks (GPON): Transmission convergence layer specification”.

STABILNOŚĆ CZĘSTOTLIWOŚCI NISKOBUDŻETOWYCH PLATFORM SDR W PROCEDURACH LOKALIZACJI OPARTYCH O METODY DOPPLEROWSKIE

FREQUENCY STABILITY OF LOW BUDGET SDR PLATFORMS IN LOCATION PROCEDURES BASED ON DOPPLER METHOD

Kacper Bednarz[1]; Jarosław Wojtuń[2];

[1] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, kacper.bednarz@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, jaroslaw.wojtun@wat.edu.pl

DOI: 10.15199/59.2023.4.56

Streszczenie: **SDF (ang. signal Doppler frequency) jest jedną z metod lokalizacji bazujących na efekcie Dopplera. Jej dokładność jest ściśle związana z dokładnością określania przesunięcia częstotliwości Dopplera. W związku z tym, w sensorach lokalizacyjnych niezbędne jest wykorzystanie wysoce stabilnych częstotliwościowo platform radia programowalnego (SDR). W tym artykule przyjrzano się dwóm szeroko dostępnym platformom SDR, które mogłyby zostać wykorzystane do budowy sensora. W oparciu o dokładność pomiaru częstotliwości chwilowej w czasie, oszacowano parametr stabilności częstotliwości każdej z badanych platform SDR.**

Abstract: **Signal Doppler frequency (SDF) is one of the methods for localization of radio emitters, which is based on the Doppler effect. Its accuracy is closely related to the accuracy of determining the Doppler frequency shift. Therefore, it is necessary to use highly frequency-stable software-defined radio (SDR) platforms in location sensors. In this paper, we explore two widely-available SDR platforms that could be used in that sensor. Based on the accuracy of instantaneous frequency measurement over time, we evaluate the frequency stability parameter of each SDR platform.**

Słowa kluczowe: **efekt Dopplera, lokalizacja, łączność bezprzewodowa, radio definiowane programowo, stabilność częstotliwości.**

Keywords: **Doppler effect, frequency stability, localization, software-defined radios, wireless communication.**

1. WSTĘP

Stabilność częstotliwości jest jednym z podstawowych parametrów określanych dla nadajników radiowych. Parametr ten jest kluczowy do prawidłowej realizacji ważnych procesów telekomunikacyjnych, takich jak synchronizacja czy wykorzystanie zasobów radiowych. W przypadku metody lokalizacji SDF [1] stabilność częstotliwości odbiornika wpływa na dokładność estymacji pozycji emitera [2].

W ramach realizowanego projektu opracowywany zostaje Dopplerowski sensor lokalizacji, który będzie oparty na radiu definiowanym programowo. Docelowo sensor zostanie zainstalowany na bezzałogowym statku powietrznym (BSP), co zapewni mu dużą mobilność, swobodę poruszania się i większy zasięg detekcji niż platforma lądowa. Celem tego artykułu jest ocena stabilności częstotliwości dostępnych, komercyjnych platform COTS (ang. Commercial off-the-shelf) SDR, które potencjalnie mogłyby zostać wykorzystane w sensorze.

Układ pozostałej części artykułu zorganizowany został w następujący sposób. Sekcja II wprowadza w problem stabilności częstotliwości. Wpływ stabilności na procedury lokalizacyjne opisano w rozdziale III. Rozdział IV zawiera opis stanowiska badawczego oraz zakres rozważanych badań. W rozdziale V przedstawiono wyniki pomiarów. Opracowanie kończy krótkie podsumowanie.

2. STABILNOŚĆ I DRYFT CZĘSTOTLIWOŚCI

Stabilność częstotliwości jest bezpośrednio związana z klasą zegara. Zazwyczaj urządzenia radiowe konstruowane są w oparciu o zegary kwarcowe. Lepszą stabilność częstotliwości zapewnia oscylator kwarcowy sterowany napięciem z kompensacją temperatury (TCVCXO) lub oscylator sterowany numerycznie (NCO). Urządzenia wyższej klasy mogą być dodatkowo stabilizowane przez wewnętrzny system synchronizacji lub zewnętrzny wzorzec częstotliwości. W tym celu często używane są odbiorniki globalnego systemu nawigacji satelitarnej (GNSS), które rozprowadzają jednolitą skalę czasu. Takie podejście jest stosowane w większości nowoczesnych stacji bazowych telefonii komórkowej w celu zapewnienia odpowiedniej synchronizacji systemu. Rubidowe wzorce częstotliwości lub cezowe zegary fontannowe są używane jako wysoce stabilne zewnętrzne źródła synchronizacji.

W ogólności, przyczyną niestabilności urządzeń jest dryft częstotliwości. To niezamierzone i ogólnie arbitralne przesunięcie oscylatora od jego częstotliwości nominalnej. Do głównych przyczyn tego zjawiska zalicza się starzenie komponentów, zmiany ich temperatury lub środowiska. Dryft częstotliwości mierzony jest w Hz/s [3]. Zakładając, iż oba urządzenia pracują w zbliżonej temperaturze, zmniejszenie poziomu dryftu, świadczyć może o większej stabilności częstotliwości. Parametr ten jest wielkością bezwymiarową, którą zwykle definiuje się jako stosunek odchylenia standardowego dryftu częstotliwości mierzonej w zadanym przedziale czasu do częstotliwości nominalnej lub nośnej, przy której dokonywany jest pomiar [2]. W zależności od przedziału czasu, dla którego określa się odchylenie dryftu, wprowadza się pojęcie stabilności krótko-, średnio- i długoterminowej.

W nadajniku radiowym dryft częstotliwości może spowodować emisję nadawanego sygnału na sąsiednim kanale, jak również problemy z jego odbiorem i jednocześnie możliwość nielegalnego zakłócania transmisji na sąsiednim kanale. Stabilność częstotliwościowa urządzeń stanowi natomiast ważny element funkcjonowania procesów synchronizacji systemów telekomunikacyjnych, a pośrednio również demodulacji i dekodowania odbieranych sygnałów.

3. STABILNOŚĆ CZĘSTOTLIWOŚCI W LOKALIZACJI

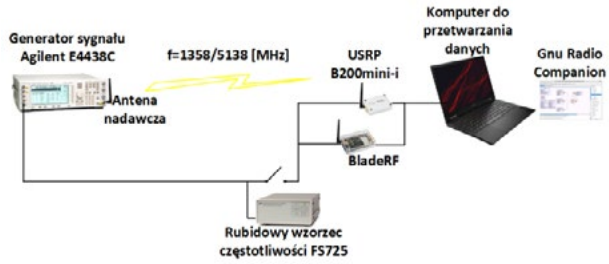
Stabilność częstotliwościowa urządzeń ma również kluczowe znaczenie w metodach lokalizacji źródeł emisji opartych o częstotliwość odebranego sygnału, np. bazująca na różnicy częstotliwości odebranego sygnału (FDOA) [4] czy SDF [1]. Metody te opierają się na efekcie Dopplera, czyli pozornej zmianie częstotliwości odbieranego sygnału, która wynika z wzajemnego ruchu obiektów – nadajnika i odbiornika. Zakres zmian przesunięcia częstotliwości Dopplera zależy przede wszystkim od częstotliwości nośnej i wypadkowej prędkości między obiektami. Dla platform lądowych i powietrznych zakres zmian częstotliwości Dopplera wynosi od kilku Hz do kilkudziesięciu kHz. Efektu Dopplera nie należy natomiast mylić z dryftem częstotliwości.

Procedura lokalizacji źródła emisji może dotyczyć dwóch ogólnych scenariuszy. Pierwszy scenariusz zakłada pozycjonowanie węzła w obrębie własnej sieci, np. ulokowanie terminala mobilnego w sieciach komórkowych. Drugi scenariusz jest typowy dla działań prowadzonych przez systemy rozpoznania i walki elektronicznej. W takim przypadku sensory lokalizacji mają za zadanie oszacować nieznaną pozycję nadajnika, co uniemożliwia synchronizację lub wymianę informacji między sensorem (odbiornikiem) a zlokalizowanym nadajnikiem. Dalsze rozważania ograniczają się do drugiego scenariusza i metody SDF.

Przy nieznanej lokalizacji emitera nie można wpłynąć na dryft częstotliwości jego zegara. Z tego względu ważne jest, aby zapewnić jak najwyższą stabilność częstotliwościową odbiornika - sensora lokalizacji. Jest ona niezbędna w trakcie trwania misji BSP, której czas dla typowych warunków operacyjnych nie przekracza godziny. W tym artykule przebadane zostały dostępne platformy COTS SDR, które można wykorzystać w sensorze lokalizacji Dopplerowskiej. Badania przeprowadzono w dwóch wariantach z wykorzystaniem zewnętrznego rubidowego wzorca częstotliwości oraz bez niego. Podobne badania dla odbiorników radiokomunikacyjnych przedstawiono w [2]. Sprzęt ten wykorzystano w badaniach empirycznych czujnika opartego na SDF zamontowanego na lądowej platformie mobilnej.

4. STANOWISKO I SCENARIUSZ POMIAROWY

Na rys. 1 przedstawiono stanowisko pomiarowe używane podczas badań stabilności częstotliwości dostępnych platform SDR. Generator sygnału Keysight (Agilent) E4438C wraz z anteną stanowi część nadawczą stanowiska pomiarowego. Dodatkowo jest on stabilizowany za pomocą rubidowego wzorca częstotliwości FS725. Po stronie odbiorczej przedstawiono platformy COTS SDR, tj. bladeRF 2.0 micro xA4 oraz B200mini-i. Są to rozwiązania niskobudżetowe, zbliżone cenowo (koszt zakupu oscyluje w granicach 1200 euro) oraz charakteryzujące się podobną masą i rozmiarami.



Rys. 1. Stanowisko pomiarowe do badań stabilności częstotliwościowej platform SDR

Metodologia pomiaru wzorowana była na przedstawionej w [2]. Badania przeprowadzono dla dwóch częstotliwości f, 1358 MHz i 5138 MHz, ze stabilizacją SDR sygnałem 10 MHz z wzorca FS725 i bez stabilizacji.

Sygnał harmoniczny o częstotliwości $f_0 = f + 10$ kHz jest podawany z generatora do anteny nadawczej. Używając SDR i oprogramowania GNU Radio Companion, próbki IQ sygnału są rejestrowane na częstotliwości f, z częstotliwością próbkowania 200 kHz dla platformy B200mini-i oraz 600 kHz dla bladeRF 2.0. Pojedynczy pomiar częstotliwości odbieranego sygnału opiera się na analizie widmowej sygnału o czasie trwania 1 s. Do analizy spektralnej wykorzystano algorytmy szybkiej transformaty Fouriera (FFT) o różnej rozdzielczości, niemniejszej niż 0,1 Hz. Pomiary prowadzono przez 45 minut. Przesunięcie 10 kHz między generatorem a SDR pozwala w analizie na ominięcie problemów wynikających z efektu DC pojawiającego się podczas przetwarzania sygnałów kwadraturowych.

5. WYNIKI BADAŃ

Po przesunięciu widma w niższy zakres częstotliwości oraz uwzględnieniu 10 kHz przesunięcia pomiędzy sygnałem nadanym a odebranym, w idealnym przypadku, częstotliwość chwilowa odebranego sygnału w paśmie podstawowym f_p powinna wynosić 0 Hz. W rzeczywistości ulega ona zmianom w czasie, z tego też względu wyznaczono jej wartość minimalną f_{min}, maksymalną f_{max} oraz średnią f_{avg}. Celem porównania stabilności platform obliczono odchylenie standardowe σ_f otrzymanych wyników oraz parametr stabilności częstotliwości s_f, wyrażony jako:

$$s_f = \frac{\sigma_f}{f_0} \tag{1}$$

gdzie f_0 – częstotliwość nośna sygnału, σ_f – odchylenie standardowe.

W tab. 1 przedstawione zostały wartości badanych parametrów dla różnych konfiguracji stanowiska pomiarowego z odłączonym po stronie odbiorczej rubidowym

wzorcem częstotliwości. Wyniki badanych parametrów z dołączonym wzorcem częstotliwości prezentuje tab. 2.

Tab. 1. Badane parametry dla wybranych platform SDR z odłączonym po stronie odbiorczej wzorcem częstotliwości

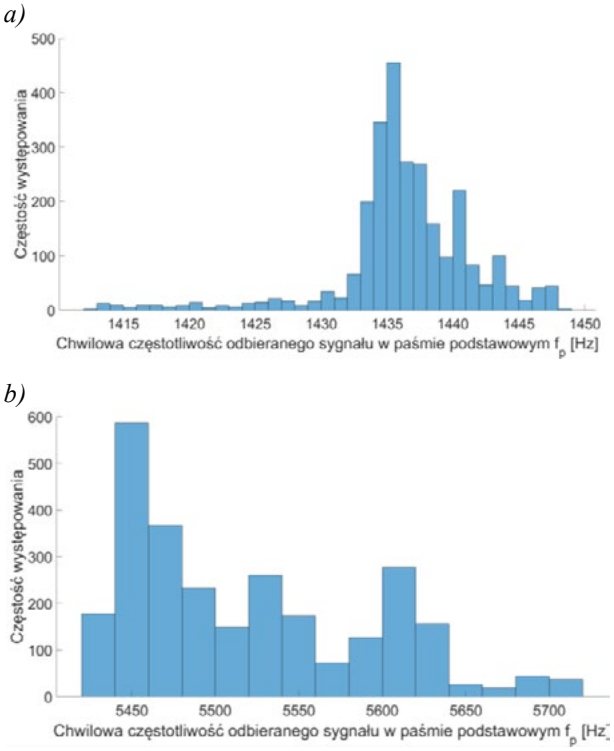
Parametr	Platforma			
	B200mini-i		bladeRF 2.0	
f [MHz]	1358	5138	1358	5138
f_{min} [Hz]	1412,80	5429,70	-70,60	-316,10
f_{max} [Hz]	1448,40	5711,70	67,30	2137,40
f_{avg} [Hz]	1436,44	5519,68	-36,46	552,71
σ_f [Hz]	5,29	71,60	34,89	955,17
s_f [-]	3,89E-09	1,39E-08	2,57E-08	1,86E-07

Tab. 2. Badane parametry dla wybranych platform SDR z dołączonym po stronie odbiorczej wzorcem częstotliwości

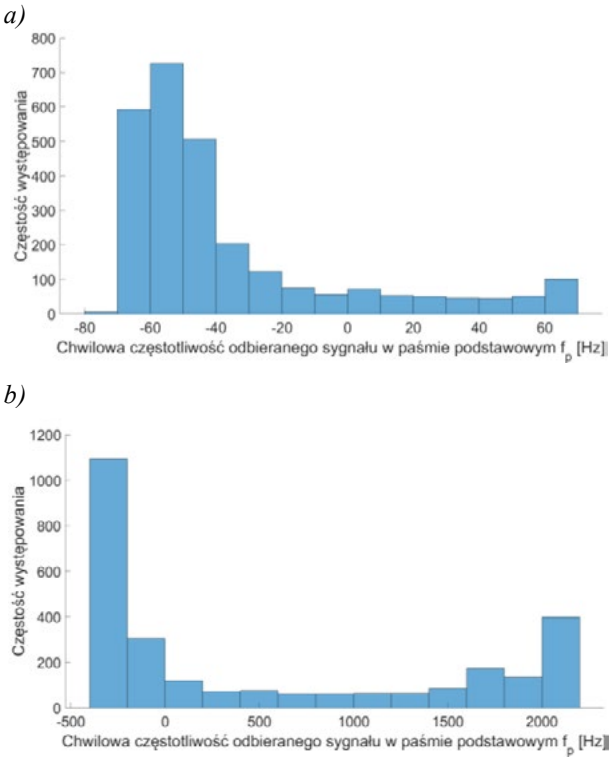
Parametr	Platforma			
	B200mini-i		bladeRF 2.0	
f [MHz]	1358	5138	1358	5138
f_{min} [Hz]	-3,70	-14,40	-0,30	-1,70
f_{max} [Hz]	3,60	14,10	-0,19	-1,64
f_{avg} [Hz]	-0,01	-0,01	-0,24	-1,67
σ_f [Hz]	1,28	5,04	4,29E-3	4,91E-3
s_f [-]	9,44E-10	9,81E-10	3,17E-12	9,56E-13

Rozkłady wartości częstotliwości chwilowej odebranego sygnału w paśmie podstawowym f_p dla dwóch zadanych wartości częstotliwości f przedstawiają rys. 2 oraz 3. Oba rysunki obrazują konfigurację stanowiska bez dołączonego zewnętrznego wzorca częstotliwości po stronie odbiorczej. Przypadek z dołączonym wzorcem częstotliwości dla platformy B200mini-i przedstawia rys. 4. Dla platformy bladeRF 90% wartości wypadało w pojedynczym przedziale <-0,240;-0,239> dla f = 1358 MHz, <-1,671;-1,670> dla f = 5138 MHz), z tego też względu nie widziano potrzeby prezentacji odpowiednich histogramów. Przebiegi częstotliwości chwilowej sygnału w funkcji czasu $f_p(t)$ bez oraz ze stabilizacją zewnętrznym rubidowym wzorcem częstotliwości obrazują kolejno rys. 5 oraz 6.

Uzyskane wyniki pozwalają na kompleksowe porównanie obu platform SDR. W przypadku pracy odbiornika bez zewnętrznego wzorca częstotliwości, widoczne jest duże odstrojenie od częstotliwości środkowej sięgające wartości tysięcy herców. Na rys. 5 zauważyć można, iż mniejsze odstrojenie posiada platforma bladeRF, jednakże po około 25 minutach pracy dla f =5138 MHz charakteryzuje się znacznym dryftem częstotliwości. B200mini-i w przypadku pracy bez wzorca posiada mniejsze odchylenie standardowe. Po jego dołączeniu sytuacja odwraca się na korzyść bladeRF. Poprawie o kilka rzędów wielkości ulega parametr stabilności częstotliwości s_f obu platform. Fluktuacje częstotliwości f_p platformy B200mini-i widoczne na rys. 6 są większe, jednakże ich średnia wartość f_{avg} z tab. 2 jest bliska zeru. BladeRF posiada niewielkie odstrojenie, które należy kompensować w przypadku chęci implementacji platformy w sensorze.

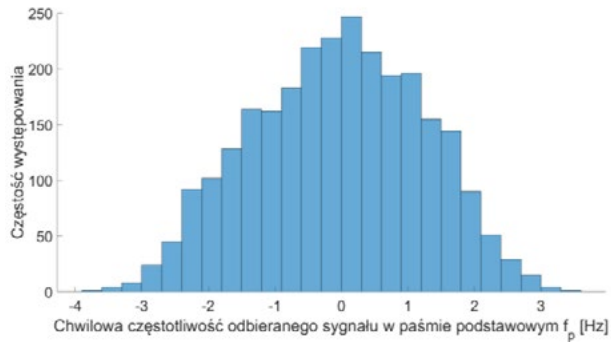


Rys. 2. Histogram f_p bez stabilizacji wzorcem platformy B200mini-i dla f: a) 1358 MHz, b) 5138 MHz

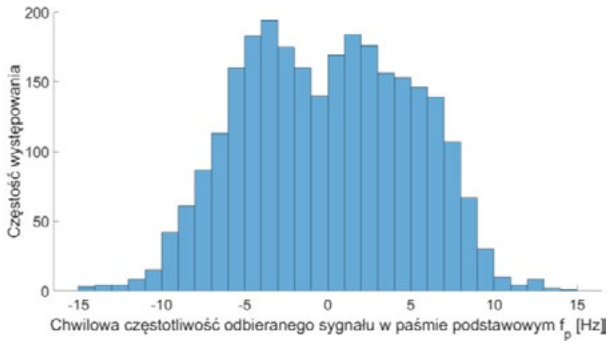


Rys. 3. Histogram f_p bez stabilizacji wzorcem platformy bladeRF 2.0 micro dla f: a) 1358 MHz, b) 5138 MHz

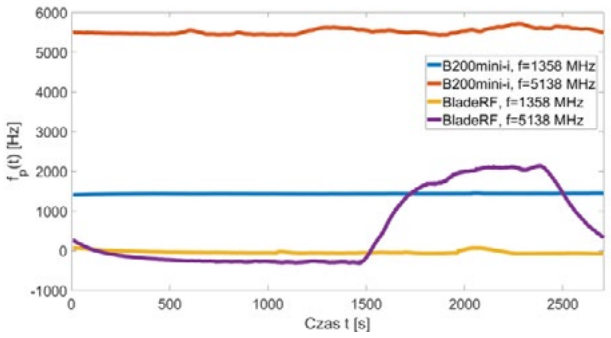
a)



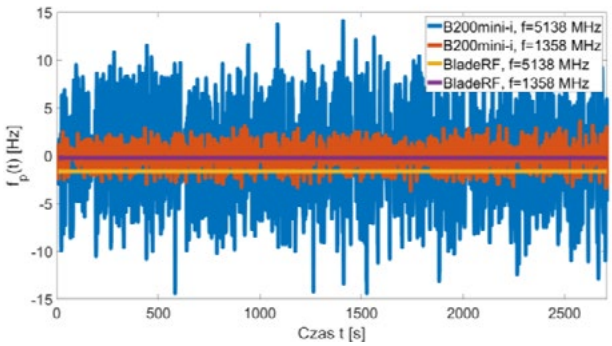
b)



Rys. 4. Histogram f_p ze stabilizacją wzorcem platformy B200mini-i dla f: a) 1358 MHz, b) 5138 MHz



Rys. 5. Przebiegi częstotliwości chwilowej sygnału w funkcji czasu $f_p(t)$ obu platform z odłączonym wzorcem częstotliwości



Rys. 6. Przebiegi częstotliwości chwilowej sygnału w funkcji czasu $f_p(t)$ obu platform z dołączonym wzorcem częstotliwości

6. PODSUMOWANIE

Celem artykułu była ocena stabilności częstotliwościowej dwóch platform COTS SDR pod kątem ich zastosowania w Dopplerowskim sensorze lokalizacji. Zaprezentowane zostało stanowisko badawcze, metodologia pomiarów oraz wyniki dla obu analizowanych platform. Badania przeprowadzono w dwóch wariantach – z rubidowym wzorcem częstotliwości, podłączonym po stronie odbiorczej i bez niego.

W przypadku pracy bez dołączonego zewnętrznego wzorca częstotliwości, bardziej stabilna okazuje się platforma B200mini-i, posiada ona jednakże większe odstrojenie od częstotliwości środkowej. Zmienia się to po dołączeniu zewnętrznego wzorca. Wartości f_p platformy B200mini-i fluktuują w zakresie kilku herców, podczas gdy odchylenie standardowe σ_f bladeRF osiąga wartości tysięcznych części herca. W przypadku drugiego SDR, posiada on jednakże wspomnianą wcześniej niezerową wartość średnią. W przyszłości planuje się przebadanie innych egzemplarzy bladeRF w celu wykluczenia możliwości występowania odstrojenia wyłącznie w analizowanej platformie.

Charakterystyczną cechą pojawiającą się podczas badań obu SDR jest również wzrost wartości odchylenia σ_f, w momencie zwiększania częstotliwości nośnej f_0, co jednak nie zawsze znajduje odzwierciedlenie w pogorszeniu parametru stabilności s_f.

Przedstawione badania wskazują na potrzebę zastosowania w projektowanym sensorze lokalizacyjnym zewnętrznego wzorca częstotliwości dołączonego do SDR. Ponieważ taki sensor będzie montowany na BSP, przebadano platformy o niewielkich rozmiarach i wadze. Analiza porównawcza pozwoliła na wybór rozwiązania charakteryzującego się najmniejszym dryftem oraz największą stabilnością częstotliwości.

LITERATURA

[1] Kelner J. M., Ziółkowski C., "Effectiveness of mobile emitter location by cooperative swarm of unmanned aerial vehicles in various environmental conditions," Sensors, vol. 20, no. 9, Art. no. 9, May 2020, doi: 10.3390/s20092575.

[2] Kelner J. M., Ziółkowski C., and Marszałek P., "Influence of the frequency stability on the emitter position in SDF method," in 2016 17th International Conference on Military Communications and Information Systems (ICMCIS), Brussels, Belgium, May 2016, pp. 1–6. doi: 10.1109/ICMCIS.2016.7496554.

[3] Lombardi M., "Time and frequency from A to Z," National Institute of Standards and Technology, May 12, 2010. https://www.nist.gov/pml/time-and-frequency-division/popular-links/time-frequency-z (accessed Mar. 20, 2023).

[4] Wang D., Zhang P., Yang Z., Wei F., and Wang C., "A novel estimator for TDOA and FDOA positioning of multiple disjoint sources in the presence of calibration emitters," IEEE Access, vol. 8, pp. 1613–1643, 2020, doi: 10.1109/ACCESS.2019.2959235.

PRECYZYJNOŚĆ POMIARU PRZEPŁYWNOŚĆI IP W SIECI LTE Z FUNKCJĄ DUAL CONNECTIVITY

MEASUREMENT ACCURACY OF IP SCHEDULED THROUGHPUT IN LTE NETWORK WITH DUAL CONNECTIVITY

Arkadiusz Zięba[1]; Jarosław Sadowski[2]; Martin Kollar[3]; Krzysztof Tatarczyk[4]

[1] Nokia Solutions and Networks, Wrocław, arkadiusz.zieba@nokia.com
[2] Gdansk University of Technology, Gdańsk, jarsadow1@pg.edu.pl
[3] Nokia Solutions and Networks, Wrocław, martin.kollar@nokia.com
[4] Nokia Solutions and Networks, Wrocław, krzysztof.tatarczyk@nokia.com

DOI: 10.15199/59.2023.4.57

Streszczenie: **Przepływność protokołu IP zdefiniowana na podstawie 3GPP TS 36.314 jest jednym z najważniejszych wskaźników Quality of Service (QoS). Jednakże specyfika sieci komórkowych w architekturze rozproszonej uniemożliwia precyzyjność jej pomiaru. Artykuł ten przedstawia porównanie klasycznej architektury Long Term Evolution (LTE) w porównaniu do systemu Dual Connectivity (DC) i skupia się na nowatorskiej metodzie rozwiązania problemu bieżącej estymacji przepływności dla terminala realizującego usługę transmisji danych jednocześnie przez dwie stacje bazowe.**

Abstract: **IP scheduled throughput defined according to 3GPP TS 36.314 is one of the most important indicators for monitoring of Quality of Service (QoS). However, the nature of networks built on a distributed architecture prevents it from being accurately measured. This paper gives an overview of the classical Long Term Evolution (LTE) architecture as opposed to Dual Connectivity (DC) system and focuses on a novel method of solving the issue of current throughput estimation for UE providing the data transmission service simultaneously through two base stations.**

Słowa kluczowe: **Agregacja Nośnych (CA), Podwójna Łączność (DC), Przepływność sieci dostępowej RAN dla protokołu IP z perspektywy terminalu, Kluczowy wskaźnik wydajności (KPI), Long Term Evolution (LTE), Jakość usługi (QoS).**

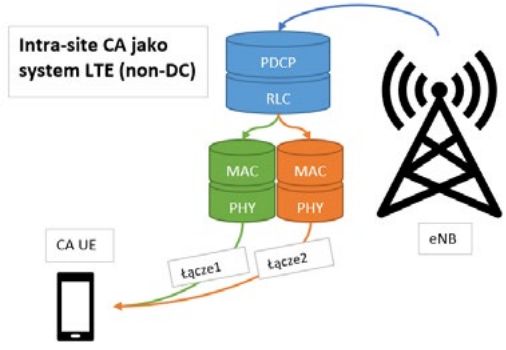
Keywords: **Carrier Aggregation (CA), Dual Connectivity (DC), IP Scheduled Throughput, Key Performance Indicator (KPI), Long Term Evolution (LTE), Quality of Service (QoS).**

1. WSTĘP

Stale rosnące zapotrzebowanie użytkowników końcowych na strumieniowanie plików wideo wysokiej jakości o dużej objętości w czasie rzeczywistym wymuszają zmiany w sieciach komórkowych nowej generacji. Ewolucja architektury sieci LTE, od systemu z wykorzystaniem pojedynczej nośnej, do systemu LTE-A Pro z wykorzystaniem dwóch nośnych poprzez funkcjonalność Intersite Carrier Aggregation, zapewnia taką możliwość. Jednak samo wykorzystanie systemów DC, które są także systemami rozproszonymi pozwalającymi uzyskać zwiększone przepływności, wiąże się z problemem pomiaru przepływności ze względu na rozłączność jednostek sterujących ruchem danych. W przypadku architektury systemów wykorzystujących jedną nośną lub dwie nośne w tej samej stacji bazowej (BTS) taki problem nie istnieje, ponieważ wszystkie niezbędne informacje do poprawnego oszacowania ruchu są dostępne w jednostce funkcjonalnej [1, 2]. Warto zaznaczyć, że przepływność IP dla operatorów sieci komórkowych jest jednym z najistotniejszych współczynników świadczenia usług. Realizacja pomiarów przepływności jest zatem obowiązkiem operatorów, a zaobserwowane błędy w realizacji pomiarów zgodnie ze standardem wymagają działań naprawczych, zaprezentowanych w niniejszym artykule.

System DC z punktu widzenia użytkownika końcowego zapewnia jednoczesną współpracę pomiędzy grupą określonych węzłów sieciowych dzięki czemu można uzyskać większe przepływności. Jednakże, procesy dotyczące poprawności transmisji danych na niższych warstwach stosu protokołów radiowych są wykonywane w każdym węźle z osobna, są więc dla siebie nieznane. Zakres zastosowania i przydatność systemów DC obejmuje sytuacje gdy aktualne wykorzystanie zasobów radiowych w danym obszarze jest niewielkie lub gdy potrzebne jest zastosowanie mechanizmów równoważenia obciążenia. Natomiast warunkiem koniecznym jest wzajemne objęcie danego obszaru zasięgiem dwóch stacji bazowych, o różnych częstotliwościach środkowych, przez co użytkownik końcowy może skorzystać z dwóch nośnych jednocześnie.

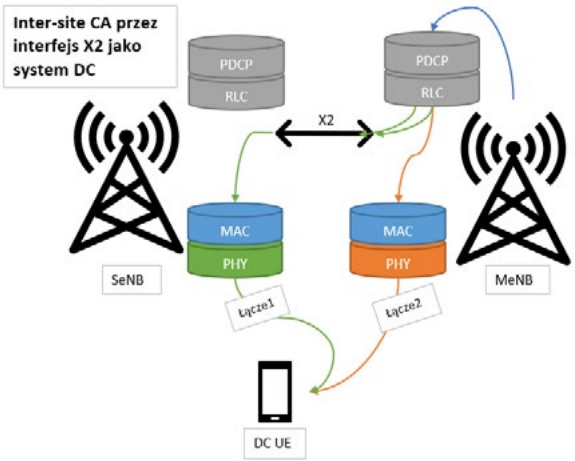
Rys.1 przedstawia klasyczny system gdzie wszystkie urządzenia sieci dostępowej RAN są fizycznie położone w tej samej stacji bazowej.



Rys. 1. Klasyczny system LTE z jedną stacją bazową i jednym użytkownikiem końcowym jednocześnie podłączonym do dwóch różnych nośnych.

Pojedynczy BTS steruje całym procesem podziału przesyłania pakietów IP poprzez różne warstwy stosu protokołów radiowych. Pakiety są kierowane do transmisji na głównej nośnej w danej stacji bazowej (PCell), na nośnej pomocniczej (SCell) lub na obu jednocześnie. Stacja bazowa jest w stanie określić poprawność transmisji dla każdego pakietu danych przez kanał radiowy na podstawie mechanizmów kontroli jakości transmisji (odpowiedź HARQ). Jeśli pakiety Packet Data Convergence Protocol (PDCP) będą wymagały ponownej transmisji na niższych warstwach ze względu na brak odpowiedzi od użytkownika końcowego lub wystąpienie błędów w kanale radiowym, to taka procedura będzie wykonywana do momentu zwolnienia bufora warstwy PDCP.

Rys.2 przedstawia system DC, gdzie użytkownik końcowy (UE) jest jednocześnie podłączony do dwóch różnych stacji bazowych, dla systemu LTE jest to Master eNB (MeNB) i Secondary eNB (SeNB) [10]. Interfejs X2 zapewnia możliwość komunikacji pomiędzy dwoma stacjami bazowymi. Każda ze stacji bazowych nie jest jednak świadoma czynności wykonywanych u swojego sąsiada. MeNB jest w stanie określić wielkość pakietów PDCP, a także czas jaki jest potrzebny na transmisję części całej sekwencji pakietów wewnątrz własnej komórki. Natomiast nie jest w stanie określić czasu jaki jest potrzebny do przesłania danych na niższych warstwach w SeNB (czas transmisji może się wydłużyć, np. ze względu na błędy w dekodowaniu wiadomości lub zmienne warunki kanału radiowego). Prostym sposobem rozwiązania problemu byłoby wykorzystanie mechanizmu odpowiedzi na każdy pakiet PDCP z osobna ze znacznikami czasu odnośnie rozpoczęcia i końca transmisji danych w SeNB. Jednakże takie podejście mogłoby przeciążyć interfejs X2, a także wydajność jednostki sterującej w MeNB, która musiałaby odpowiednio powiązywać odebrane informacje z własnymi na niższych warstwach (ze względu na rozdział ruchu). Wynika to z faktu, że im większy jest ruch IP przesyłany z MeNB do SeNB tym więcej potwierdzeń będzie wymagane dla każdego z pakietów.



Rys. 2. System DC dla LTE z dwoma stacjami bazowymi i jednym użytkownikiem końcowym jednocześnie podłączonym do dwóch różnych nośnych w dwóch różnych stacjach.

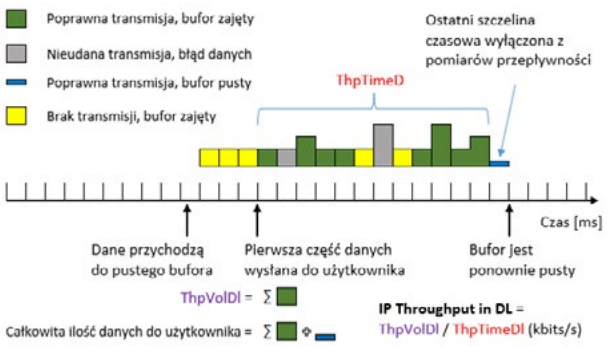
Takie podejście mogłoby również ograniczyć liczbę pakietów danych możliwych do przesłania pomiędzy stacjami bazowymi na rzecz utrzymania mechanizmu kontroli (potwierdzeń), jednak celowość tego działania byłaby dyskusyjna, ponieważ świadomie zostałby ograniczony zysk z przepływności.

W kolejnej części referatu przedstawiono koncept standardowej metody pomiaru przepływności w systemie E-UTRAN zdefiniowanej w 3GPP TS 36.314. Następnie przedstawiono nową metodę do oszacowywania przepływności w systemach DC, a na końcu zaprezentowano wyniki porównawcze z zastosowania obu metod pomiarowych.

2. STANDARDOWA METODA POMIARU PRZEPŁYWNOŚCI IP W E-UTRAN

W dokumentacji 3GPP TS 32.451 zostało zdefiniowane wymaganie pomiaru przepływności z perspektywy użytkownika, które bazuje na ilości danych z poziomu protokołu IP (PDCP) oraz na czasie trwania pojedynczej transmisji, zwanej burst'em. Bezczynność bufora danych (T_Idle, tj. czas, w którym w buforze nie oczekują dane do wysłania) określa granicę pomiędzy dwoma burst'ami do siebie przyległymi, a także nie jest brana pod uwagę przy pomiarze przepływności [8].

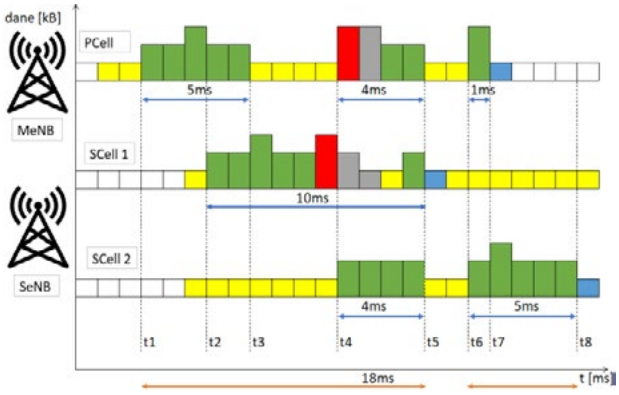
Sposób standardowej metody pomiaru przepływności IP (IP Scheduled Throughput) jest zdefiniowany w dokumentacji 3GPP TS 36.314 [6]. Na Rys. 3 przedstawiono podstawowy koncept zasady pomiaru przepływności IP na przykładzie pojedynczego burst'a danych w kierunku do użytkownika dla pojedynczej komórki. Ilość poprawnie przesłanych danych z warstwy IP jest określona jako ThpVolDl, natomiast czas trwania pojedynczego burst'a jako ThpTimeDl. Warto zauważyć, że został wprowadzony podział pomiędzy stanami zajętości bufora danych warstwy PDCP, co skutkowało tym, że czas oczekiwania na rozpoczęcie transmisji dla danego użytkownika nie będzie zliczany. Ponadto, ostatnie szczeliny czasowe, rozumiane jako Transmission Time Intervals (TTI) dla systemu LTE, również zostały pominięte podczas zliczania ThpTimeDl. Powodem takiej decyzji jest fakt, że ostatnie szczeliny nigdy nie są w pełni wypełnione danymi użytkownika. Wynika to z określonej metodyki wyboru wielkości bloku danych, Transport Block (TB), w warstwie MAC w pojedynczej szczelinie zależnej od aktualnych warunków w kanale radiowym. To powoduje, że ostatni TTI wypełniany jest sztucznymi danymi określanymi jako padding volume. Zawarcie takich danych przy metodzie pomiaru skutkowałoby zmianą samej precyzji pomiaru, która mogłaby się okazać niedoszacowana [6, 7, 9].



Rys. 3. Zasada działania standardowej metody pomiaru przepływności IP w kierunku do użytkownika.

3. PRECYZJA POMIARU PRZEPŁYWNOŚCI IP STANDARDOWEJ METODY W SYSTEMACH DC

Na Rys. 4. przedstawiono przykładowe rozważania odnośnie błędu pomiaru przepływności dla systemu DC z wykorzystaniem standardowej metody. Pakiety danych IP zostały podzielone pomiędzy trzy komórki, główną PCell w MeNB i dwie pomocnicze SCell 1 i SCell 2 znajdujące się w SeNB. Poprawne transmisje danych do użytkownika oznaczono kolorem zielonym, błędne i wymagające ponownej transmisji kolorem czerwonym, powtórzone transmisje kolorem szarym, transmisje w ostatniej szczelinie czasowej kolorem niebieskim, zajętość bufora kolorem żółtym, natomiast kolorem białym został oznaczony pusty bufor. Dla osi x pojedyncza jednostka czasu to 1 ms, dla osi y ilość danych to od 1 kB do 3 kB, w zależności od wielkości pakietu danych.



Rys. 4. Problemy z precyzją pomiaru przepływności IP w kierunku do użytkownika w systemach DC.

W celu zrozumienia problemu należy rozważyć następujące zdarzenia, pamiętając, że MeNB nie posiada żadnych informacji odnośnie czasu transmisji w SeNB, a także opóźnienia na interfejsie X2.

Pierwsze dotyczy sytuacji kiedy ilość danych PDCP jest równa sumie pakietów danych skierowanych do komórki głównej i pomocniczych, natomiast czas zliczany jest wyłącznie z perspektywy PCell i bazuje na znacznikach czasu od t1 do t3 i od t4 do t5 i od t6 do t7. Tego rodzaju rozwiązanie zdecydowanie zawyży estymatę przepływności (Opcja 1 w Tab.1).

Drugie dotyczy sytuacji kiedy ilość danych jest równa sumie pakietów danych skierowanych do komórki głównej i pomocniczych, natomiast czas jest mierzony z perspektywy PCell oraz SCell 1 i bazuje na znacznikach czasu od t1 do t5 i od t6 do t7. Pomimo, że zliczany jest czas przerwy w transmisji w PCelli, to jednak w tym samym momencie odbywa się transmisja na SCell 1, o czym MeNB posiada informacje, ze względu na wcześniej wysłane dane do SeNB. Niestety nie jest zliczany czas od t7 do t8 gdzie odbywa się transmisja na SCell 2. Tego rodzaju rozwiązanie zarówno może zawyżyć estymatę przepływności (Opcja 2 w Tab.1) jak i obniżyć jeśli przerwy w transmisji będą dłuższe niż czas trwania transmisji w obu komórkach.

Trzeci scenariusz obejmuje sytuację gdy zliczana jest ilość danych wyłącznie w PCell, natomiast przy pomiarze czasu wliczany jest także dodatkowo czas przerwy w transmisji danych w obu stacjach bazowych w tym samym momencie, od t1 do t7. Takie rozwiązanie zaniży estymatę przepływności (Opcja 3 w Tab.1).

Tab. 1. Błędy pomiaru standardowej metody dla systemu DC.

Punkt odniesienia	*Ilość danych [kB]*	*Czas transmisji [ms]*	*Przepływność [Mbps]*
PCell	21	10	16,80
SCell 1	16	10	12,80
SCell 2	19	9	16,89
Opcja 1	56	10	44,80
Opcja 2	56	14	32,00
Opcja 3	21	16	10,50
Użytkownik	56	18	24,89

Tab.1 podsumowuje wcześniej omówione wyniki. Z perspektywy użytkownika dla zadanego przykładu przepływność wyniosła 24,89 Mbps. Żadne z powyższych rozważań nie pozwoliłyby obliczyć poprawnej przepływności IP z punktu widzenia użytkownika dla standardowej metody w systemach DC.

4. SPOSOBY ROZWIĄZANIA PROBLEMU

W pierwszym rozdziale wspomniano oczywisty sposób rozwiązania problemu precyzji pomiaru dla standardowej metody w systemach DC, na podstawie mechanizmu potwierdzeń ze znacznikami czasu. Dla takiego podejścia każda komórka pomocnicza musiałaby przesyłać znacznik czasu rozpoczęcia i zakończenia danego pakietu PDCP. Taka metoda wiązałaby się z obciążeniem interfejsu X2 danymi nieistotnymi z punktu widzenia poprawności samej transmisji. SeNB musiałaby raportować informacje dotyczące poprawności przesyłania pakietów dla każdego użytkownika z osobna, korzystającego z agregacji nośnych. Oznacza to, że kosztem przesyłania danych do użytkownika interfejs rezerwowałby część swojej przepustowości dla danych informacyjnych, nieistotnych dla użytkownika, jednak istotnych dla operatora. Zależność obciążeniowa byłaby wprost proporcjonalna do ilości danych użytkownika przesyłanych przez interfejs. Jednakże, poważniejszym problemem byłoby trzymanie w pamięci informacji odnośnie poprawności transmisji w SeNB i powiązywanie tych danych dla każdego burst'a i każdego użytkownika, co znacznie obniżyłoby wydajność całej stacji bazowej.

Innym rozwiązaniem nieobniżającym wydajności BTS'a i interfejsu X2 byłoby zastosowanie aproksymacji przepływności w danej komórce pomocniczej i przesłanie takiej informacji co określony interwał czasowy do komórki głównej. Dzięki takiemu podejściu obciążenie interfejsu byłoby zawsze stale określone, w przeciwieństwie do wcześniej przedstawionego rozwiązania. Ponadto, jednostka sterująca ruchem w MeNB nie musiałaby powiązywać informacji odnośnie rozpoczęcia i zakończenie transmisji na różnych nośnych dla danego użytkownika, ponieważ SeNB informowałoby o swojej aktualnej przepustowości dla każdego użytkownika z osobna.

Dla takiego rozwiązania sugerowane jest także wykorzystanie już istniejącego mechanizmu, który periodycznie raportuje wartości z jednego BTS'a do drugiego. Downlink Data Delivery Status (DDDS) zdefiniowany w 3GPP TS 36.425 zapewnia taką możliwość [11].

5. NOWA METODA POMIARU PRZEPŁYWNOŚCI IP W SYSTEMACH DC

W celu osiągnięcia sposobu precyzyjnego szacowania przepływności IP dla użytkownika końcowego zarówno dla systemów klasycznych jak i DC niezbędne jest zdefiniowanie podstawowych założeń.

Po pierwsze, rozpoczęcie zliczania czasu transmisji pojedynczego burst'a jest wyłącznie zależne od bufora po stronie MeNB. Jest to także zgodne z wrażeniami jakich doświadcza użytkownik końcowy, który połączony jest z komórką główną, podejmującą decyzję o aktywacji komórek pomocniczych w zależności od zapotrzebowania na wielkość ruchu i samej zdolności terminala do świadczenia określonych funkcji transmisyjnych. Kolejno należy założyć, że znacznik czasu otrzymania pierwszego pakietu RLC PDU po stronie SeNB jest nieistotny.

Po drugie, należy zdefiniować bufor nadrzędny (główny) w stosunku do każdego z buforów poszczególnych stacji bazowych MeNB i SeNB. Bufor główny będzie zajęty jeśli jeden z buforów MeNB lub SeNB będzie zajęty.

Po trzecie, czas zajętości bufora w SeNB dla każdego pakietu RLC PDU$_i$ należy określić jako:

$$T'_{RLC\ PDU_i} = T_{X2_RLC\ PDU_i} + T_{RLC\ PDU_i} \quad (1)$$

$T_{X2_RLC\ PDU_i}$ jest to czas transmisji danego pakietu przez interfejs X2. Natomiast $T_{RLC\ PDU_i}$ jest to czas przez który dany pakiet RLC PDU$_i$ znajduje się w buforze SeNB, obliczany w następujący sposób:

$$T_{RLC\ PDU_i} = \frac{ilość\ danych\ RLC\ PDU_i}{Przepływność\ IP\ SeNB} \quad (2)$$

Przepływność IP SeNB jest mierzona na podstawie standardowej metody (3GPP TS 36.314) wyłącznie w samym SeNB dla otrzymanych pakietów RLC PDU od MeNB.

Po czwarte, wartości uśrednione czasu opóźnienia transmisji przez interfejs X2 oraz przepływności IP SeNB zostaną wykorzystane do obliczenia (1) i (2). Zostanie również zdefiniowany parametr sterujący częstotliwością ich wysyłania z SeNB do MeNB. W zależności od intensywności ruchu pomiędzy dwoma stacjami bazowymi, będzie można dostosować parametr do takiej wartości, która nie doprowadzi do przeciążenia interfejsu X2, co skutkowałoby obniżeniem zysku przepływności przez podział ruchu między różne stacje bazowe.

Powyższe założenia pozwolą uzyskać wynik pomiaru przepływności IP z perspektywy użytkownika końcowego (24,89 Mbps) jak wskazano w Tab.1. Czas o wartości 18ms zawiera przerwę w transmisji w MeNB, ale ze względu na zajęty bufor w SeNB nadal jest brany pod uwagę. Natomiast, przerwa, kiedy obydwa bufory w MeNB i SeNB są puste, nie jest istotna (wlicza się tutaj też ostatnia szczelina czasowa dla transmisji na SCell 1). Ilość danych jest sumaryczną wartością wielkości pakietów PDCP otrzymanych w MeNB od sieci rdzeniowej [5].

6. WYNIKI POMIARÓW PRECYZJI PRZEPŁYWNOŚĆI IP DLA RUCHU TCP

Poprawność precyzji pomiaru przeprowadzono na oprogramowaniu sprzętu firmy Nokia, dla którego nowa i standardowa metoda została wdrożona w postaci liczników sieciowych, raportujących informacje w określonych odstępach czasowych. Liczniki dostępne są z poziomu MeNB, która steruje ruchem danych. Wartości liczników określono w Tab. 2. za pomocą ThpVolDl oraz ThpTimeDl, opisujących kolejno całkowity wolumen danych i czas transmisji z punktu widzenia użytkownika.

Scenariusz testowy dotyczył transmisji danych o małej wielkości (25MB) w ustalonych odstępach czasowych (10s) do użytkownika za pomocą protokołu TCP. Charakter tego testu ma na celu naśladować pobieranie niewielkich plików, np. podczas przeglądania stron internetowych. Należy także zwrócić uwagę, że charakterystyka ruchu TCP, wykorzystująca odpowiedzi zwrotne do serwera wysyłającego informacje źródłowe, nie pozwala na uzyskanie tak wyraźnego wzrostu przepływności, aniżeli w przypadku ruchu z wykorzystaniem protokołu UDP. Przez co aktywacja komórek pomocniczych nie następuje w początkowej fazie transmisji.

Jednym z celów tego testu było osiągnięcie minimalnego pułapu danych, po którym następuje aktywacja komórek pomocniczych dla funkcjonalności agregacji pasm w celu osiągnięcia zwiększonych przepływności [3, 4]. Dlatego tylko około 30% danych podczas trwania pojedynczego burst'a jest kierowana do komórek pomocniczych, natomiast pozostała reszta jest przesyłana w komórce głównej. Test miał również wykorzystać rodzaj konfiguracji, który pozwoliłby osiągnąć zbliżone wyniki pod kątem przepływności dla klasycznego systemu LTE i systemu DC, co pozwoliłoby policzyć błąd względny dla każdej z metod z osobna. Ze względu na różnice wydajności i w sposobie działania obu technik transmisji LTE i DC, można spodziewać się niewielkich odchyleń wartości przepływności, pomimo zbliżonej konfiguracji testu.

Tab. 2. Pomiary przepływności przy użyciu standardowej i nowej metody w systemach LTE i DC dla ruchu TCP o małych burst'ach i konfiguracji CA 3CC.

	LTE		DC	
	Standardowa	Nowa	Standardowa	Nowa
ThpVolDl [Mbit]	6060	6154	4095	5881
ThpTimeDl [ms]	24050	25473	24014	24077
Przepływność [Mbps]	251,97	241,59	170,53	244,27
Błąd względny standardowej metody [%]	100*(170,53 - 251,97) / 251,97 = -32,32			
Błąd względny nowej metody [%]	100*(244,27 – 241,59) / 241,59 = 1,11			

Tab. 2. przedstawia wyniki testów przeprowadzonych dla konfiguracji składających się z 3 komórek o tej samej szerokości pasma (20 MHz) dla pojedynczego użytkownika. Wykorzystany rodzaj modulacji to 256QAM, natomiast liczba warstw MIMO została ustawiona na 2 dla każdej z komórek. Bezwzględny numer kanału częstotliwości radiowej (E-ARFCN) dla PCell to 1100 (1970 MHz – 1990 MHz), SCell 1 to 66600 (2116,4 – 2136,4 MHz) i SCell 2 to 67000 (2156,4 – 2176,4 MHz).

Test był wykonywany w tym samym czasie w systemie LTE dla obu metod, gdzie uzyskano wyniki przepływności dla standardowej i nowej metody kolejno 251,97 i 241,59 Mbps. W tym scenariuszu stacja bazowa posiadała wszystkie informacje, aby poprawnie policzyć przepływność. Natomiast w systemie DC, dla standardowej metody, MeNB już takich informacji nie posiadała, stąd wynik stanowczo zaniżony 170,53 Mbps. Błąd względny obu metod porównuje więc dwa różne scenariusze testowe, które co prawda nie różnią się ilością danych jaką należy wysłać do użytkownika ani szerokością całkowitego pasma częstotliwości, lecz różnią się sposobem dostarczenia tych informacji. Metoda standardowa nie uwzględnia transmisji przez nośne na drugiej stacji bazowej dla systemu DC, natomiast uwzględnia ruch przez wszystkie nośne (nawet za pomocą różnych anten) na tej samej stacji bazowej dla systemu LTE. Nowa metoda w stosunku do standardowej pozwala policzyć przepływność dla transmisji danych przez nośne na drugiej stacji bazowej z perspektywy głównej stacji bazowej.

Błąd względny uzyskany dla nowej metody wyniósł 1,1%, natomiast dla standardowej metody aż -32,32%. Powodem takiej różnicy dla standardowej metody jest zliczanie pakietów PDCP tylko przesyłanych przez PCell. Zbliżone wielkości zmierzonego czasu świadczą o tym, że ruch kierowany do komórek pomocniczych był kierowany w większości podczas przesyłania danych w buforze w komórce głównej, co uwarunkowane jest kontrolowanym środowiskiem testowym z jednym użytkownikiem. Warto zwrócić uwagę, że w przypadku testów w sieciach klienckich, gdzie do jednej komórki może być jednocześnie podłączonych nawet 200 lub więcej użytkowników, taka zależność nie będzie istnieć. Zależność ta dotyczy mechanizmu kolejkującego użytkowników w odpowiednich szczelinach czasowych (pojedynczy użytkownik nie musi mieć przydzielanych zasobów w następujących po sobie szczelinach czasowych) w zależności od priorytetów zdefiniowanych dla określonych charakterystyk ruchu, a także na podstawie wybranych ustawień operatora oraz warunków w kanale radiowym.

Motywacją testu było wykazanie błędu precyzji pomiaru standardowej metody w systemach DC oraz poprawność nowej metody w klasycznych systemach LTE (bez wykorzystania DC), a także w systemach DC dla pojedynczego użytkownika.

7. PODSUMOWANIE

Artykuł ten dowiódł następujących tez. Po pierwsze, standardowa metoda nie jest w stanie precyzyjnie oszacować pomiaru przepływności IP dla użytkownika końcowego dla systemów DC (skonstruowanych z elementami sieci dostępowej w różnych miejscach geograficznych). Po drugie wykazała poprawność szacowania przepływności IP dla zaproponowanej nowej metody zarówno w systemach klasycznych, takich jak LTE (ruch odbywa się przez pojedynczy punkt zbiorczy sieci dostępowej), jak i w systemach DC.

Przyszłe badania powinny skupić się na weryfikacji mechanizmu raportującego średnie opóźnienie oraz średnią przepływność pomocniczej stacji bazowej (SeNB) oraz wpływu zmiennych warunków w kanale radiowym na nowo zdefiniowane współczynniki.

LITERATURA

[1] E. Dahlman, S. Parkvall and J. Skold. 2013. „4G: LTE/LTE-Advanced for Mobile Broadband (Second Edition)". Academic Press. 544.

[2] Harri Holma, Antti Toskala and Takehiro Nakamura. 2019. „5G Technology: 3GPP New Radio". Nokia Bell Labs, Finland.

[3] Anders Hedlund and Irina Cotanis. 2014. „An Introduction to Carrier Aggregation Testing". Ascom Network Testing.

[4] Jari Salo and Eduardo Zacarías B. 2017. „Analysis of LTE Radio Load and User Throughput". International Journal of Computer Networks & Communications (IJCNC) Vol.9, No. 6.

[5] M. Kollar, M. Tomala and A. Zięba. 2022. „Evaluation of dl ip scheduled throughput for inter enb carrier aggregation". Patent numer US11523407B2.

[6] 3GPP Standard TS 36.314. 2020. „Evolved Universal Terrestrial Radio Access (E-UTRA); Layer 2 - Measurements". Version 16.0.0, Release 16.

[7] 3GPP Standard TS 32.450. 2020. „Key Performance Indicators (KPI) for Evolved Universal Terrestrial Radio Access Network (E-UTRAN): Definitions". Version 16.0.0 Release 16.

[8] 3GPP Standard TS 32.451. 2019. „Key Performance Indicators (KPI) for Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Requirements". Version 16.0.0 Release 16.

[9] 3GPP Standard TS 32.425. 2020. „Performance Management (PM); Performance measurements; Evolved Universal Terrestrial Radio Access Network (E-UTRAN)". Version 17.0.0 Release 17.

[10] 3GPP Standard TS 23.501. 2021. „System architecture for the 5G System (5GS); Stage 2". Version 17.0.0 Release 17.

[11] 3GPP Standard TS 36.425. 2022. „X2 interface user plane protocol". Version 17.0.0 Release 17.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

Dynamiczne dostosowywanie wag łączy dostępowych w multi-homingu

Dynamic adaptation of upstream connection weights in multi-homing

Piotr Jurkiewicz[1]; Michał Rzepka[1]; Piotr Jaglarz[1]; Jerzy Domżał[1]; Robert Wójcik[1];

[1] link2net s.c., kontakt@link2net.pl

DOI: 10.15199/59.2023.4.58

Streszczenie: W artykule przedstawiono rezultaty badań nad mechanizmem pozwalającym na jednoczesny adaptacyjny dostęp do Internetu przy użyciu wielu sieci dostępowych w warunkach zmieniającej się ich przepustowości. Zaproponowano dwa algorytmy dostosowywania wag łączy dostępowych do zmieniającej się przepustowości, pierwszy bazujący na obserwacji średniej szybkości pobierania na istniejących połączeniach i drugi oparty na monitorowaniu czasu odpowiedzi na pakiety ping. Z badań wynika, że pozwalają one na znaczne zniwelowanie pogorszenia jakości obsługi wynikającego z zmieniającej się przepustowości łączy. Najlepiej wypada algorytm ping.

Abstract: The paper presents the results of research on mechanisms of dynamic adaptation of upstream connection weights in multi-homing. Two algorithms were proposed, the first one based on observation of the download speed of existing connections and the second one based on the monitoring of ping probe packets RTT. Research shows that their usage significantly eliminates QoS deterioration resulting from variable bandwidth of upstream connections. The ping algorithm achieves the best results.

Słowa kluczowe: ruter, przepływy, tablica przepływów, inżynieria ruchu, multi-homing

Keywords: router, flows, flow table, traffic engineering, multi-homing

1. WSTĘP

Najpowszechniej znanym typem urządzeń sieciowych są domowe rutery IP. Przekazują one pakiety pomiędzy urządzeniami w sieci lokalnej użytkownika (LAN), a siecią dostawcy usług (WAN), która z kolei jest połączona z Internetem. Zazwyczaj sieć lokalna podłączona jest przez ruter tylko do jednego dostawcy usług internetowych (ISP).

Podłączenie do wielu dostawców mogłoby zwiększyć niezawodność, ponieważ gdy jedno z łączy przestawałoby działać, pakiety mogłoby być przesyłane przez łącza pozostałych operatorów. Połączenie hosta lub sieci końcowej do więcej niż jednej sieci dostępowej nazywane jest *multi-homingiem*. Sieci te nie muszą być tego samego typu. Na przykład, jedna z sieci dostępowych może być kablową siecią Ethernet dostarczaną przez ISP, podczas gdy druga może być siecią komórkową, podłączoną poprzez modem bezprzewodowy. W takim przypadku sieć komórkowa może służyć za łącze awaryjne, używane gdy sieć przewodowa przestanie działać.

Niektóre z obecnie dostępnych na rynku ruterów klasy domowej posiadają możliwość podpięcia wielu łączy dostępowych, jednakże transmisja w danej chwili może odbywać się tylko przy użyciu jednego z nich. Drugi i kolejne interfejsy są wykorzystywane jedynie w przypadku awarii tego pierwszego.

Połączenie poprzez wiele sieci dostępowych mogłoby być jednak także wykorzystywane do zwiększenia wydajności. Używanie wielu łączy jednocześnie do przesyłania danych pozwoliłoby na zwiększenie osiąganej przepustowości. Dodatkowo, w zależności od obciążenia sieci dostępowej lub mocy sygnału, bardziej efektywne w danej chwili mogłoby być używanie konkretnej sieci dostępowej.

W artykule przedstawiono rezultaty badań symulacyjnych nad mechanizmem pozwalającym na równoległy i adaptacyjny dostęp do Internetu przy użyciu wielu sieci dostępowych. Zakłada się dostęp z dwóch źródeł, np. drogą radiową i kablową. Zakłada się, że w sieci lokalnej ruch będzie obsługiwany za pomocą przepływów, które będą rozróżnialne za pomocą pięciu pól nagłówka pakietów obsługiwanych w ruterze (adres źródłowy IP, adres docelowy IP, port źródłowy, port docelowy, identyfikator protokołu transportowego). Istotą mechanizmu ma być adaptacyjne wybieranie łącza dostępowego w zależności od obciążenia sieci tak, aby by możliwe było uzyskanie zrównoleglonej transmisji i utrzymanie jak największej sumarycznej wydajności w warunkach zmieniającej się w czasie przepustowości łączy dostępowych.

2. POWIĄZANE PRACE

W przypadku dużych ruterów korporacyjnych i operatorskich naturalnym jest używanie wielu sieci do przesyłania ruchu jednocześnie. Jest to możliwe dzięki protokołom rutingu działającym w tych sieciach, np. BGP, które znajdują najlepsze trasy do poszczególnych destynacji. Jednakże rutery klasy domowej z powodów praktycznych i bezpieczeństwa nigdy nie są podłączone do protokołów rutingu działających w sieciach operatorskich. Oznacza to, że nie jest możliwe zastosowanie w nich mechanizmów znanych z sieci operatorskich.

Większość dostępnych na rynku ruterów IP do zastosowania domowego i biurowego nie zapewnia żadnego rodzaju *multi-homingu*. Wyjątkiem są niektóre rutery firmy ASUS, np. model GT-AX11000 [4]. Pozwalają one na skonfigurowanie drugiej sieci WAN, która może być połączona zarówno poprzez kabel Ethernet, jak i bezprzewodowy modem USB. Mogą one działać w dwóch trybach: *load-balance* i *fail over*. W trybie *load-balance* użytkownik ustawia stosunek w jakim ruter będzie przydzielał nowe

połączenia na obie sieci WAN. Stosunek ten jest konfigurowany przez użytkownika i nie zmienia się w czasie. W trybie *fail over* używana jest tylko jedna sieć WAN. Ruter okresowo przesyła zapytania DNS przez tą sieć i gdy wykryje, że przestała ona działać, przełącza się na drugą.

Istnieje także pakiet dla otwartoźródłowego oprogramowania OpenWRT [3] przeznaczonego dla ruterów wielu producentów, który zapewnia podobne funkcjonalności, o nazwie mwan3 [1]. Podobnie jak oprogramowanie firmy ASUS, pozwala on na statyczne równoważenie ruchu pomiędzy wiele sieci oraz pracę w trybie *fail over*. Pakiet pozwala także na tworzenie bardziej zaawansowanych konfiguracji, w których sieć WAN wybierana jest w zależności od źródłowego lub docelowego adresu IP, portu źródłowego lub docelowego oraz typu protokołu warstwy transportowej (*policy-based routing*). Jednak podobnie jak w przypadku oprogramowania ASUS, stosunek w jakim dzielony jest ruch jest stały i nie zmienia się w czasie.

Nie znaleźliśmy natomiast żadnych artykułów naukowych w których byłby poruszany problem adaptacyjnego dostosowywania wag do zmieniającej się w czasie przepustowości łączy w odniesieniu do zastosowania w przypadku *multi-homingu*.

3. ŚRODOWISKO SYMULACYJNE

Badania zostały przeprowadzone w zmodyfikowanym środowisku Mininet [6]. W sieci znajdował się host umieszczony za węzłem emulującym ruter z NAT. Ruter był podłączony jedynym lub dwoma łączami dostępowymi (*upstream*) (w zależności od scenariusza) do kolejnych ruterów (bez NAT) za którymi był umieszczony węzeł stanowiący serwer. Przepustowość łączy w sieci, poza łączami *upstream*, była nieograniczona.

Na serwerze był uruchomiony serwer HTTP nginx [2]. Klient, przy użyciu programu `curl` wysyłał do niego żądania HTTP GET. Wielkość pliku ściąganego w ramach żądania była zmienna, zadana rozkładem logarytmicznie normalnym o parametrach $mu = 10$ i $sigma = 2$ [7]. Dawało to średnią wielkość żądania równą: 162754, 7 bajtów. Odpowiada to rozkładowi i średniej wielkości żądań HTTP w Internecie [5].

Częstotliwość generowania żądań została dobrana tak, aby uzyskać żądany ruch na średnim na poziomie 90 Mbps. Wszystkie zmienne losowe były generowane przy użyciu generatora pseudolosowego z sterowanym *seedem*, dzięki czemu sekwencje żądań były powtarzalne pomiędzy symulacjami.

Przeanalizowano następujące scenariusze:

- 1x100 – ruter podłączony jest jednym łączem *upstream* o stałej przepustowości 100 Mbps,
- 1x50 – ruter podłączony jest jednym łączem *upstream* o stałej przepustowości 50 Mbps,
- 2x50 – ruter podłączony jest dwoma łączami *upstream* o stałej przepustowości 50 Mbps każde,
- 2x50d – ruter podłączony jest dwoma łączami *upstream* o zmiennej przepustowości.

W przypadku 2x50d (dynamicznej przepustowości) przepustowość łączy *upstream* w czasie trwania symulacji była zmienna w zakresie 5-95 Mbps, lecz sumarycznie przepustowość obu łączy wynosiła cały czas 100 Mbps. Celem tego scenariusza było zbadanie algorytmów adaptacyjnego dostosowywania wag, według których dzielone są nowe przepływy na dostępne łącza *upstream* w sytuacji zmiennej ich przepustowości (spowodowanej np. zmiennym obciążeniem w sieciach różnych operatorów lub poruszaniem się rutera w przestrzeni pomiędzy stacjami bazowymi różnych operatorów). Zmienność ta także była sterowana kontrolowanym *seedem*.

4. ALGORYTMY

W scenariuszu dynamicznej przepustowości łączy (2x50d) opracowano i zbadano następujące algorytmy sterujące wagami łączy *upstream* w ruterze NAT:

- 2x50d – brak dostosowywania wag, wagi cały czas wynoszą 50/50,
- 2x50d_as_bw – stosunek wag wynosi cały czas tyle samo, co stosunek przepustowości łączy (jako scenariusz referencyjny, nie jest to możliwe do osiągnięcia w rzeczywistości),
- 2x50d_conntrack – stosunek wag obliczany na podstawie monitoringu tablic `conntrack` w ruterze,
- 2x50d_ping – stosunek wag obliczany na podstawie opóźnień i strat pakietów próbnych wysyłanych na łącza *upstream*.

4.1 Algorytm conntrack

Co 1 sekundę monitorowana jest tablica przepływów w ruterze i obliczana średnia szybkość pobierania wszystkich aktywnych połączeń na poszczególnych łączach *upstream* w trakcie ostatniej sekundy. Następnie stosunek wag tych łączy określany jest jako stosunek średnich szybkości tych przepływów. Algorytm badano z parametrem $ewm = 0, 1$, który określał zastosowaną średnią ruchomą (*exponential weighted moving average*) do wygładzania zmiany stosunku łączy: stanowiła ona $0, 9$ stosunku w poprzedniej sekundzie i $0, 1$ stosunku wynikającego ze średnich szybkości przepływów na łączach w obecnej sekundzie. Analizowano także wariant z parametrem $top = 10$, który oznaczał że w obliczeniach uwzględniano tylko 10 najszybszych przepływów na danych łączu, a nie wszystkie.

4.2 Algorytm ping

W momencie pojawienia się pierwszego pakietu nowego przepływu (ale nie częściej niż co 2 sekundy) wysyłany jest pakiet ICMP ping. W symulacjach serwerem docelowym był serwer HTTP, w rzeczywistym urządzeniu mogą to być serwery dostawcy urządzenia. Następnie stosunek wag łączy ustalany jest jako stosunek odwrotności pierwiastków czasów odpowiedzi na pakiety ping wysłane przy użyciu poszczególnych łączy *upstream*. W przypadku straty pakietu przyjmuje on jako czas odpowiedzi 1s.

5. PARAMETRY SYMULACJI

Dla każdego zbadanego przypadku, przeprowadzono symulacje z 10 różnymi wartościami *seeda* sterującego klientem HTTP oraz z 10 różnymi wartościami *seeda* sterującego dynamiczną przepustowością łączy. Dla każdej pary *seedów* symulację powtarzano 10 razy. Daje to sumaryczną liczbę 1000 symulacji dla każdego ze scenariuszy. W trakcie każdej symulacji wygenerowano 10000 żądań HTTP, z odstępami pomiędzy żądaniami wynikającymi z ruchu oferowanego na poziomie 90 Mbps. Symulacja trwała do momentu kiedy zakończone zostało pobieranie plików ze

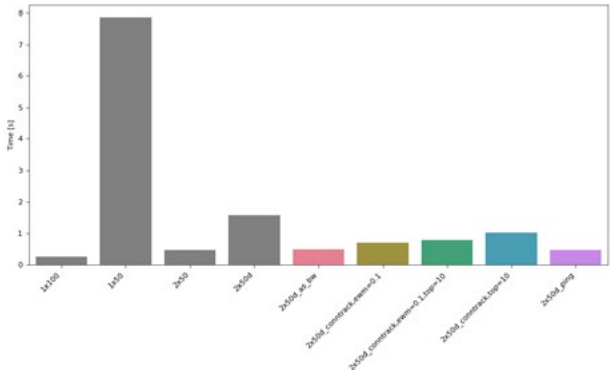
wszystkich żądań. Oznacza to że dla każdego scenariusza wygenerowano 10 000 000 żądań HTTP.

Mierzone były następujące parametry:

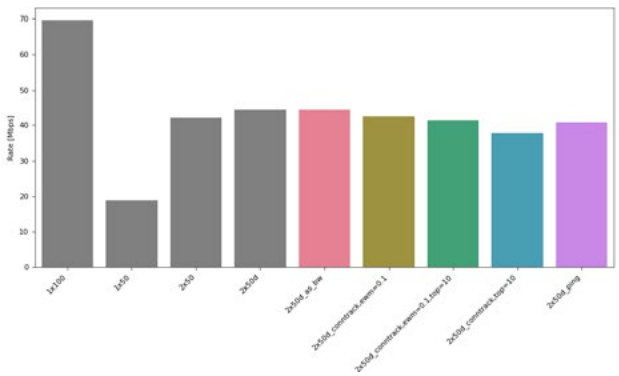
- *flow completion time (FCT)* – czas ukończenia żądania, zdefiniowany jak czas od momentu rozpoczęcia do zakończenia pobierania pliku,
- *flow rate* – średnia szybkość żądania, zdefiniowana jako wielkość pobieranego pliku podzielona przez czas jego pobierania,
- *simrate* – średnia ilość ruchu przesyłanego w jednostce czasu w czasie symulacji, zdefiniowana jako liczba bajtów pobrana w jednostce czasu poprzez wszystkie łącza *upstream*,
- *simtime* – całkowity czas symulacji, czyli czas od momentu rozpoczęcia symulacji do momentu zakończenia pobierania ostatniego pliku.

6. WYNIKI

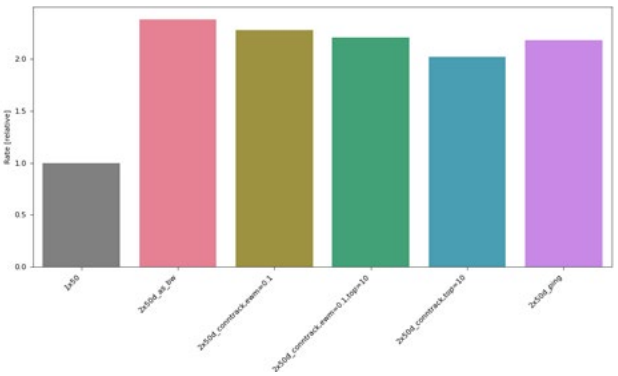
Na wykresach poniżej zaprezentowano średnie wartości wymienionych powyżej parametrów dla czterech scenariuszy testowych (szare słupki). Dodatkowo, dla scenariusza 2x50d zaprezentowano zmierzone wartości parametrów przy użyciu różnych algorytmów dynamicznego sterowania wagami (kolorowe słupki).



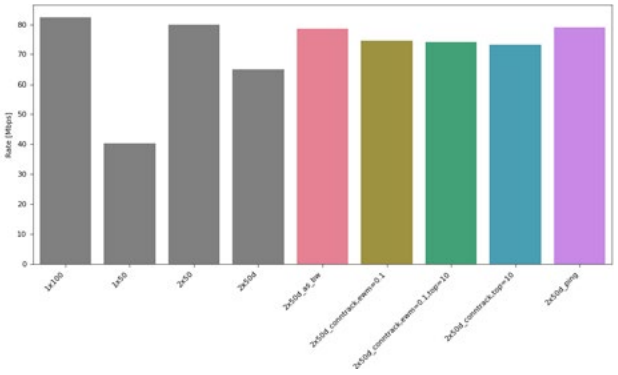
Rysunek 1: Średni czas ukończenia żądania (*mean flow completion time*)



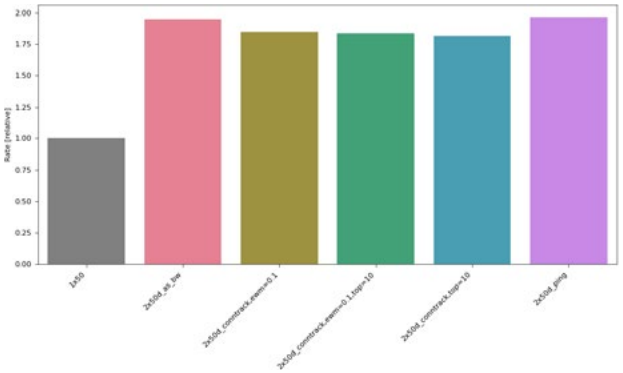
Rysunek 2: Średnia szybkość żądania (*mean flow rate*)



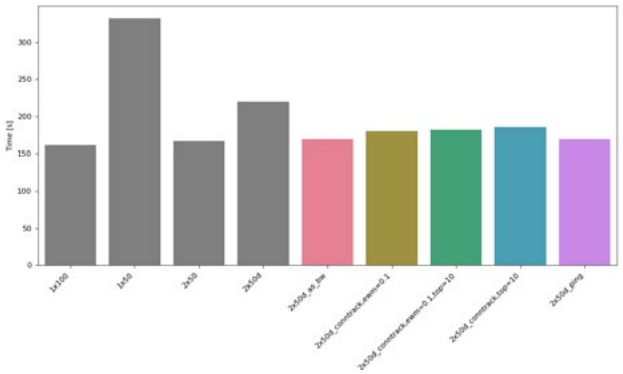
Rysunek 3: Średnia szybkość żądania (*mean flow rate*) w stosunku do scenariusza 1x50



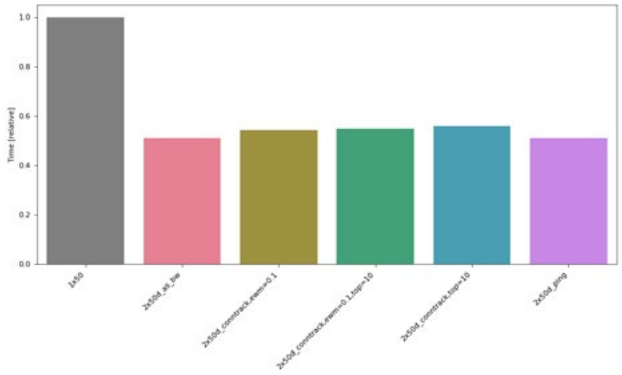
Rysunek 4: Średnia szybkość przesyłu danych (*mean simrate*)



Rysunek 5: Średnia szybkość przesyłu danych (*mean simrate*) w stosunku do scenariusza 1x50



Rysunek 6: Średni czas zrealizowania wszystkich żądań (*mean simtime*)



Rysunek 7: Średni czas zrealizowania wszystkich żądań (*mean simtime*) w stosunku do scenariusza 1x50

Wyniki zaprezentowane na Rysunkach 1-7 pokazują kolejno: średni czas ukończenia przepływu (Rys. 1), średnią prędkość przepływu (Rys. 2 i 3), średnią prędkość całkowitego przesyłu danych na łączu (Rys. 4 i 5) oraz średni czas transmisji danych (Rys. 6 i 7).

Na Rysunku 1 widać, że najkrótszy średni czas ukończenia żądania występuje w scenariuszu gdy mamy do dyspozycji jedno łącze o sumarycznej przepustowości 100 Mbps. Najdłuższy czas z kolei występuje przy użyciu jednego łącza o przepustowości 50 Mbps. Warto zauważyć, że czas ukończenia żądania nie skaluje się liniowo wraz z przepustowością łącza. Wynika to z faktu, że przy ruchu oferowanym na poziomie 90 Mbps, pojedyncze łącze 50 Mbps jest znacznie przeciążone. Straty pakietów powodują potrzebę oczekiwania na retransmisje utraconych pakietów, co znacznie zwiększa czas realizacji żądań.

W przypadku dwóch łączy o stałej przepustowości 50 Mbps czas jest nieznacznie wyższy niż w przypadku jednego łącza 100 Mbps. Natomiast gdy przepustowość łączy zaczyna dynamicznie fluktuować, czas ten wzrasta ponad dwukrotnie, pomimo tego, że sumaryczna przepustowość obu łączy jest stała. Wynika to z faktu, że nowe żądania przydzielane są na oba łącza cały czas w stosunku 50/50, więc w sytuacji gdy jedno z łączy ma chwilowo niską przepustowość także staje się ono znacznie przeciążone, podczas gdy drugie łącze jest niedociążone.

Zastosowanie algorytmów adaptacyjnie sterujących wagami pozwala na obniżenie czasu. Najlepsze rezultaty daje użycie wag równych chwilowej przepustowości łączy (teoretyczny scenariusz referencyjny) oraz algorytm oparty na pomiarze RTT (ping).

W przypadku średniej szybkości żądania (Rysunek 2 i 3) wyniki są podobne, chociaż różnice nie są już tak znaczące. Na Rysunku 4 i 5 zaprezentowano z kolei sumaryczną średnią szybkość przesyłu danych. Jest ona zbliżone przy zastosowaniu jednego łącza 100 Mbps i dwóch łączy 50 Mbps o stałych przepustowościach. Spada jednak gdy przepustowość łączy zmienia się dynamicznie. Zastosowanie algorytmów adaptacyjnego sterowania wagami pozwala na osiągnięcie przy zmieniającej się przepustowości łączy sumarycznej średniej szybkości przesyłu danych na poziomie podobnym do tego, gdyby przepustowości łączy były stałe. Najlepiej wypada tutaj algorytm ping. Podobnie sytuacja ma się w przypadku średniego czasu zrealizowania wszystkich żądań (Rysunek 6 i 7).

7. PODSUMOWANIE

W artykule zaprezentowano i przebadano dwa algorytmy dynamicznego dostosowywania wag łączy dostępowych w przypadku ich zmiennej przepustowości. Pierwszy algorytm bazuje na obserwacji tablicy połączeń i obliczaniu średniej szybkości pobierania na poszczególnych łączach dostępowych, a drugi na wysyłaniu pakietów próbkujących i monitorowaniu czasu odpowiedzi. Dodatkowo w celu referencyjnym przebadany został algorytm powielający chwilowe przepustowości łączy jako ich wagi.

Po pierwsze należy zauważyć, że użycie dwóch łączy powoduje uzyskanie gorszych parametrów jakości obsługi niż jednego łącza równego ich sumarycznej przepustowości. Jest to zgodne z oczekiwaniami i wynika z tzw. prawa wiązki. Z badań wynika, że parametry ulegają jednak dalszemu pogorszeniu w sytuacji gdy przepustowość poszczególnych łączy zmienia się w czasie (pomimo, że jej suma pozostaje cały czas stała). Użycie zaprezentowanych algorytmów dynamicznego dostosowywania wag łączy pozwala na częściowe zniwelowania tego pogorszenia. Najlepiej wypada algorytm ping, oparty na wysyłaniu pakietów próbkujących i monitorowaniu czasu ich odpowiedzi.

PODZIĘKOWANIA

Artykuł powstał w ramach projektu "Innowacyjny ruter o zwiększonej szybkości transmisji danych" o numerze RPMP.01.02.01-12-0612/18 finansowanego w ramach Regionalnego Programu Operacyjnego Województwa Małopolskiego na lata 2014-2020.

LITERATURA

[1] "mwan3 (Multi WAN load balancing/failover)," https://openwrt.org/docs/guide-user/network/wan/multiwan/mwan3, accessed: 2023-05-26.

[2] "NGINX: Advanced Load Balancer, Web Server, & Reverse Proxy," https://www.nginx.com/, accessed: 2023-05-26.

[3] "OpenWRT project," https://openwrt.org/, accessed: 2023-05-26.

[4] "[Wireless Router] Dual WAN introduction and setup - Failover and Load Balance," https://www.asus.com/support/FAQ/1011719/, accessed: 2023-05-26.

[5] P. Jurkiewicz, G. Rzym, and P. Boryło, "Flow length and size distributions in campus internet traffic," *Computer Communications*, vol. 167, pp. 15–30, 2021.

[6] B. Lantz, B. Heller, and N. McKeown, "A network in a laptop: Rapid prototyping for software-defined networks," in *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*, ser. Hotnets-IX. New York, NY, USA: Association for Computing Machinery, 2010.

[7] P. Jurkiewicz, "flow-models: A framework for analysis and modeling of IP network flows," *SoftwareX*, vol. 17, p. 100929, 2022.

OTOCZENIE RADIOWE LINII KOLEJOWYCH W KONTEKŚCIE ZAKŁÓCEŃ ŁĄCZNOŚCI W SYSTEMIE GSM-R

RADIO ENVIRONMENT OF RAILWAY LINES IN THE PRESENCE OF GSM-R COMMUNICATION INTERFERENCE

Marek Sumiła[1]

[1] Instytut Kolejnictwa, Warszawa, msumila@ikolej.pl

DOI: 10.15199/59.2023.4.59

Streszczenie: **W artykule przedstawiono wyniki analiz i badań radiowych sąsiedztwa publicznej infrastruktury radiokomunikacyjnej i obszarów kolejowych, dla których przewidywana jest zabudowa i uruchomienie systemu radiołączności kolejowej GSM-R. W kolejnych punktach przedstawiono istotę i zakres wdrażania GSM-R w Polsce, ogólny stan wiedzy w zakresie występowania zakłóceń w łączności użytkowników sieci GSM-R, wyniki analizy otoczenia radiowego dla wybranego fragmentu linii kolejowej i wreszcie wyniki pomiarów radiowych wykonanych przez Instytut Kolejnictwa.**

Abstract: **The article presents results of analyses and measurements of the public radio communication infrastructure adjacent to railway areas for which it is planned to build and launch a GSM-R railway radio communication system. The following sections present the purpose and scope of the implementation of the GSM-R system in Poland, the general state of knowledge concerning the occurrence of interference in the communication of GSM-R network users, the results of an analysis of the radio environment for a selected part of the railway line and finally the results of radio measurements carried out by the Railway Research Institute.**

Słowa kluczowe: **łączność kolejowa, GSM-R, zakłócenia, bezpieczeństwo**

Keywords: **railway communication, GSM-R, interference, safety**

1. WSTĘP

Wskazany w tytule artykułu system GSM-R (ang. *GSM for Railway*) jest obowiązującym od ponad dwudziestu lat rozwiązaniem technicznym na potrzeby radiołączności kolejowej tor – pojazd. System dostarcza dwie podstawowe usługi: głosową łączność radiową oraz radiową wymianę danych na potrzeby sterowania.

Na potrzeby tego systemu zostało wydzielone pasmo radiowe UIC (fr. *Union Internationale des Chemins de fer*) w zakresach 876–880 MHz (ang. *Upink*) i 921–925 MHz (ang. *Downlink*) [4]. Co istotne, GSM-R został wskazany w Technicznych Specyfikacjach Interoperacyjności (ang. *Technical Specification for Interoperability*) systemu kolei w obszarze Sterowanie [10] i jest obligatoryjny na nowo budowanych, remontowanych i modernizowanych liniach kolejowych w UE.

GSM-R wpisany jest w idee realizacji pociągów dużych prędkości i pośrednio odpowiedzialny jest za bezpieczeństwo jazdy pociągu poprzez komunikację pokładowego i naziemnego podsystemu odpowiedzialnego za sterowanie ruchem za pośrednictwem zautomatyzowanego systemu ETCS (ang. *European Train Control System*) poziom 2 i 3.

Polska na mocy przyjętych porozumień, w ramach Unii Europejskiej oraz wyznaczonych korytarzy transportowych [9], realizuje program wdrożenia systemu GSM-R na blisko 15 tys. km linii kolejowych [6].

Równocześnie, od blisko dekady, raportowane są przez zarządców infrastruktury i przewoźników w Europie przypadki występujących zakłóceń w komunikacji poprzez sieć GSM-R urządzeń pokładowych (radiotelefonów oraz modemów EDOR (ang. *ETCS Data Only Radio*) do transmisji danych dla systemu ETCS) w określonych lokalizacjach na terenach kolejowych [11]. Przeprowadzone na szeroką skalę działania angażujące organizacje standaryzujące, organizacje kolejowe, zarządców infrastruktury kolejowej oraz ośrodki i laboratoria badawcze doprowadziły do ustalenia głównych przyczyn występujących interferencji. Są to zakłócenia intermodulacyjne oraz blokowanie odbiorników na skutek silnych sygnałów pozapasmowych. Wyniki analiz można znaleźć w raportach komitetu technicznego CEPT (fr. *Conférence Européenne des administrations des Postes et des Télécommunications*) oraz dokumentach wydanych przez UIC [1]-[3]. Wskazane opracowania dość dobrze opisują naturę zakłóceń i wskazują sposoby ich ograniczenia.

Pomimo dość dobrze rozpoznanego zjawiska zakłóceń wielu operatorów sieci GSM-R w Europie boryka się z problemem występowania zakłóceń w części lokalizacji swoich sieci. Stan ten jest związany ze złożonością działań, które muszą być podjęte do jego rozwiązania oraz odpowiedzialnością podmiotów zaangażowanych (operator sieci GSM-R, operatorzy publiczni, regulator krajowy). Polska, implementując na swoim terenie interoperacyjny system radiołączności kolejowej, musi również podjąć działania prowadzące do ograniczenia skutków możliwych interferencji.

W dalszej części tego artykułu zostanie przedstawiony stan wiedzy w zakresie występujących zakłóceń sieci GSM-R, uwarunkowania formalno-prawne oraz przykładowy fragment linii kolejowej w Polsce, który ma być objęty zasięgiem budowanej sieci GSM-R. Właśnie dla tego fragmentu przedstawiono wyniki analizy otoczenia nadajników sieci publicznych operatorów i wyniki badań radiowych wykonanych w 2022 r.

2. STAN WIEDZY I PRZEGLĄD LITERATUROWY

Decyzje o przyjęciu standardu zapewniającego możliwość interoperacyjności komunikacji radiowej w ramach europejskiego systemu kolei przyjęto jeszcze w latach dziewięćdziesiątych XX wieku. Wybrano wówczas technologię bazującą na publicznym standardzie GSM drugiej generacji, który wydawał się wówczas najlepszym wśród dostępnych standardów, rozwiązaniem. Był on już wówczas dojrzałym rozwiązaniem, posiadającym wsparcie organizacji standaryzujących oraz sprawdzonym przez publicznych operatorów. GSM-R był rozszerzoną wersją GSM 2G; dostosowaną do wymogów kolejowych poprzez dostępność wielu dedykowanych i oczekiwanych przez zarządców i przewoźników usług dodatkowych. Zasadniczy opis obecnej wersji standardu wskazano w specyfikacjach EIRENE [4][5].

Niespodziewanie, następstwem rozwoju technologicznego sieci publicznych (3G, później 4G) na obszarach przyległych do infrastruktury kolejowej stały się zakłócenia w komunikacji radiowej realizowane poprzez sieć kolejową. Raport CEPT ECC 162 [2] dowodzi, że bezpośrednio narażone na zakłócenia są terminale użytkowników sieci GSM-R pracujące w sąsiedztwie nadajników operatorów świadczących usługi publiczne (925 ÷ 960 MHz). Istniała w tym względzie uzasadniona obawa, że źródłem zakłóceń są emisje pozapasmowe od tych właśnie nadajników. W następstwie przeprowadzonych badań laboratoryjnych (DBNetz) odrzucono ta tezę, gdyż stosowane przez operatorów sieci publicznych urządzenia były zgodne z wymaganiami normatywnymi, a ich parametry nie przekraczały dopuszczalnych poziomów emisji sygnałów poza granicami pasma.

Nie zmienia to faktu, że odbiorniki systemu GSM-R są mało odporne na silne sygnały z pozapasmowe i silne sygnały wywołujące zjawiska intermodulacyjne w obwodach wejściowych tych odbiorników. Raport [3] wskazuje, że zjawisko intermodulacji może zachodzić na skutek sygnałów pochodzących od nadajników pracujących z technologiami wąskopasmowymi, jak i szerokopasmowych. Wyniki przeprowadzonych analiz dowodzą, że szczególnie groźne są intermodulacje od nadajników szerokopasmowych pracujących w paśmie przylegającym do pasma UIC, czyli 925 ÷ 935 MHz. Ocenia się, że wypływ na występowanie zakłóceń mogą mieć nadajniki usytuowane w odległościach kilkuset metrów (do 700 m / 1 km) od linii kolejowej, zatem najczęściej poza obszarem należącym do zarządcy kolejowego.

Podsumowaniem prac jest raport ECC 229 [3], w którym podsumowano zebrano wyniki prac i zaproponowano działania mitygujące zakłócenia w sieci GSM-R.

3. UWARUNKOWANIA FORMALNO-TECHNICZNE

Przedstawione w punkcie 2 informacje wskazują, iż w odniesieniu do możliwego wystąpienia zjawiska zakłóceń sieci GSM-R, istotne jest ograniczenie oddziaływania nadajników operatorów publicznych w obszarach przylegających do infrastruktury kolejowej pracujących w paśmie 900 MHz. Działania te leżą bezpośrednio poza zasięgiem możliwości operatorów sieci GSM-R, gdyż każdy operator publiczny, na mocy uzyskanej licencji oraz zezwoleń środowiskowych, ma prawo do emisji sygnałów zapewniających określony poziom dostępności i jakości oferowanych usług. W tym kontekście prawo to dotyczy również abonentów podróżujących transportem kolejowym. Oznacza to, że nie jest w interesie operatorów publicznych ograniczenie dostępu do usług również w obszarach kolejowych. Wobec powyższego, zgodnie z zaleceniami raportu ECC 229 [3] operator sieci GSM-R powinien monitorować stan swojej sieci i poszukiwać drogi dialogu z operatorami sieci publicznych, których infrastruktura radiowa prowadzi do powstawania zakłóceń. Innym rozwiązaniem jest zaangażowanie strony trzeciej, którą może być instytucja mająca kompetencje i mandat do rozstrzygania powstających sporów.

Odpowiedzią na zaistniałą sytuację, Komisja Europejska wraz z UIC, dążąc do zachowania spójności wymagań w zakresie wybranego systemu interoperacyjnej radiołączności kolejowej, wprowadziła w aktualizacji TSI CCS z 2016 r. [10] wymóg zastosowania specyfikacji technicznej ETSI TS 102 933-1 [14] dla wszystkich nowo certyfikowanych pojazdów oraz dodatkowy parametr 4.2.17 wprowadzający wymóg dodatkowych testów radiowych RSC (ang. *Radio System Compatibility*) dla pojazdów poruszających się po określonej infrastrukturze kolejowej. Dla wykonawców radiowej infrastruktury GSM-R wprowadzono wymóg potwierdzenia parametrów QoS (ang. *Quality of Service*) z zastosowaniem dwóch różnych podsystemów CCO (ang. *on-board Control-Command and signalling*), czyli przejazdu dwóch certyfikowanych pojazdów wyposażonych w pokładowy system ETCS oraz radio kabinowe GSM-R. Dla zarządców posiadających już certyfikowaną infrastrukturę radiokomunikacyjną żadnych nowych wymagań nie wprowadzono.

4. OTOCZENIE RADIOWE LINII KOLEJOWYCH W POLSCE

Polska jest aktualnie w trakcie realizacji dużego projektu budowy sieci GSM-R. Działania te są następstwem przyjętego przez Polski Rząd Krajowego Planu Wdrażania ERTMS (ang. *European Rail Traffic Management System*) i dotyczy ok. 15 tys. km linii kolejowych [6]. W drugiej dekadzie XXI wieku zrealizowano i zcertyfikowano już ok. 2 tys. km linii kolejowych.

Komercyjnie w Polsce system GSM-R jest użytkowany zaledwie na dwóch liniach kolejowych, tj. na linii kolejowej nr 9 relacji Warszawa Wsch. – Gdynia Gł. (transmisja danych na potrzeby sterowania w systemie ETCS) oraz na linii Pomorskiej Kolei Metropolitalnej w okolicach Gdańska na potrzeby komunikacji głosowej.

Eksploatowany na wskazanych liniach system GSM-R stał się źródłem pierwszych doświadczeń dla polskich przewoźników, jak i operatora sieci (PKP PLK S.A.). Jednym z tych doświadczeń były zaobserwowane utrudnienia w komunikacji radiowej w okolicach przystanku osobowego Gdynia Wzgórze św. Maksymiliana.

Jest to pierwszy potwierdzony przypadek występowania zakłóceń GSM-R w Polsce. Źródłem zakłóceń jest nadajnik BTS (ang. *Base Transceiver Station*) publicznego operatora sieci komórkowej, którego jedna z anten jest skierowana w kierunku linii kolejowej w okolicy wspomnianego przystanku kolejowego.

Autor niniejszego artykułu w swoich wcześniejszych publikacjach (m.in. [12][13]) posiłkując się wiedzą europejskich publikacji rozważał różne aspekty zakłóceń systemu GSM-R w Polsce. Zidentyfikowane, źródło (BTS) i jego operator potwierdzają zakładane wówczas scenariusze.

Uzyskane wyniki stały się również motywacją do szerszego spojrzenia na otoczenie radiowe fragmentu linii kolejowej w okolicach Gdyni. W tym celu wykorzystano informacje udostępnione przez UKE (pl. *Urząd Komunikacji Elektronicznej*) oraz narzędzie dostępne na stronie http://beta.btsearch.pl/ w celu identyfikacji obiektów radiokomunikacyjnych znajdujących się pomiędzy stacją Gdynia Gł., a przystankiem osobowym Wejherowo stanowiącego przedłużenie linii Warszawa – Gdynia Gł. w kierunku Kołobrzegu oraz wykonanie pomiarów poziomu sygnału radiowego nadajników pracujących w paśmie 925 ÷ 935 MHz jako paśmie bezpośrednio przylegającym do pasma dedykowanego dla GSM-R.

4.1. Usytuowanie publicznych BTS w otoczeniu linii kolejowych

Przyjęty do analizy fragment otoczenia linii kolejowej nr 202 na odcinku 23 km nie posiada aktualnie aktywnej infrastruktury GSM-R. Fragment ten jednak został wskazany przez operatora sieci do przyszłej zabudowy urządzeń i będzie włączony obszarowo do sieci GSM-R.

Rysunek 1 przedstawia mapę obejmującą okolice Gdyni w kierunku Rumii do Wejherowa. Pogrubioną szarą linią zaznaczono położenie torów głównych szlakowych na wskazanym terenie. Dla przejrzystości wizualizacji na mapie włączono wyłącznie obiekty radiowe należące do operatora posiadającego koncesję Prezesa UKE do nadawania w paśmie 925 ÷ 930 MHz. Lista obiektów radiowych wszystkich operatorów sieci komórkowych posiadających nadajniki na objętym analizą obszarze jest większa.

Rys. 1. Mapa okolic Gdynia – Wejherowo z zaznaczoną linii kolejowej szarym kolorem

Załączona mapa dobrze obrazuje sąsiedztwo nadajników na obszarach przyległych do terenów kolejowych.

Raporty [2][3][7][8][11] wskazują na obiekty radiowe na dystansie do 700 m od obszaru kolejowego. Na ocenianym fragmencie zidentyfikowano 30 obiektów radiowych na tym dystansie i dodatkowo 5 obiektów, których odległość nie przekracza 1 km. Średnia odległość tych obiektów od torów wynosi 268 m (dla dyst. 700 m), a najbliższy BTS był usytuowany w odległości ok. 30 m (odchylenie standardowe 187,02 m). Odległość średnia pomiędzy kolejnymi obiektami radiowymi wzdłuż linii kolejowej wynosiła ok. 767 m.

Oczywistym jest, że przedstawiona analiza wymaga uzupełnienia o dodatkowe informacje, które mogą mieć wpływ na wielkość narażeń. Zaliczyć do nich należy przeszkody terenowe pomiędzy antenami, a obszarem ocenianym, jak również szereg parametrów technicznych opisujących każdy włączonych do analizy BTS-ów. Mowa tu o wysokości zawieszenia anten, ich pochyleniu (ang. *Tilt*), azymucie oraz mocy emitowanej EIRP (ang. *Effective Isotropical Radiated Power*). Wskazane parametry nie są dostępne w zastosowanym narzędziu internetowym, a ich pozyskanie wymaga czasu i może być utrudnione. Właściwym autorowi wydało się przeprowadzanie pomiarów radiowych na obszarze kolejowym w miejscach, w których odległość do publicznego BTS jest orientacyjnie najmniejsza.

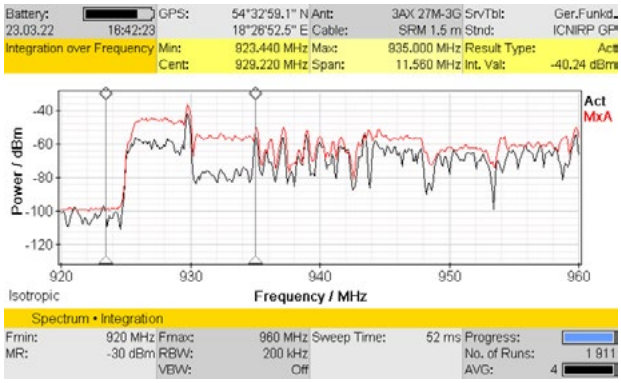
4.2. Wyniki pomiarów radiowych dla wybranego fragmentu linii kolejowej

Zaplanowane w 2022 roku pomiary radiowe dla wytypowanego odcinka linii kolejowej miały stanowić potwierdzenie istotności wstępnych analiz prowadzonych na podstawie rozważań teoretycznych.

Jako kryterium oceny uzyskanych empirycznie wyników przyjęto specyfikację techniczną ETSI TS 102 933-1 [14], która określa m.in. wymagane parametry dla odbiornika radiowego, których spełnienie wpłynie na ograniczenie negatywnego wpływu sygnałów zakłócających i wyższą odporność na interferencje. Jednym z tych parametrów jest wartość dopuszczalnego sygnału pozapasmowego, na który radiotelefon powinien być odporny. Wartość tę ustalono na poziomie -40 dBm.

Pomiary radiowe wykonano wzorcowanym przyrządem firmy NARDA typu SRM-3006 w paśmie 920 ÷ 960 MHz. Pomiary wykonywano punktowo w pasie kolejowym, możliwie blisko zewnętrznej krawędzi toru, na wysokości ok. 3 m (w specyfikacjach EIRENE SRS [4] zasadniczo przyjmuje się wysokość 4 m jako wysokość zamontowania anteny taborowej w pojeździe, jednak ze względu na potrzebę uwzględnienia również użytkowników radiotelefonów przenośnych (ok. 1,5 ÷ 2,0 m) wysokość tę zredukowano do 3 m). Dodatkowo, ze względu na występujące zjawisko fast fading'ów mierzono oprócz wartości chwilowej również wartość średnią dla pasma radiowego.

Przykładowy wynik pomiaru w jednej z lokalizacji przedstawiono na rysunku 2.



Rys. 2. Przykładowy wynik pomiaru aktywnych nadajników w paśmie 900 MHz

Na powyższym wykresie kolorem czarnym zaznaczono wynik pomiaru chwilowego, kolorem czerwonym uśrednione wartości sygnałów wykonanych w punkcie pomiarowym (ponad 1900 przebiegów).

Uzyskane wyniki dla tej konkretnej lokalizacji, w paśmie bezpośrednio przylegającym do pasma UIC, dowodzą aktywności nadajnika pracującego z technologią szerokopasmową o średniej wartości sygnału w bloku 5 MHz zbliżonym do dopuszczalnej granicy -40 dBm oraz z technologią 2G na skraju pasma (930 MHz) i sile sygnału wyższej od przyjętego poziomu -40 dBm.

W sumie pomiary wykonano w 13 lokalizacjach, dla których dystans do BTS publicznego operatora był najmniejszy. Wśród nich dla siedmiu lokalizacji wartości sygnałów pozapasmowych oscylowały w granicach -40 ÷ -60 dBm.

5. PODSUMOWANIE

System GSM-R jest jednym z zasadniczych elementów koniecznych do zapewnienia interoperacyjności systemu kolei w Europie. Jego zastosowanie ma wpływ na komunikację głosową maszynista – obsługa naziemna, ale również na efektywne sterowanie pociągiem poprzez wymianę danych w systemie ETCS. Podjęta jeszcze w latach dziewięćdziesiątych ubiegłego wieku decyzja o wyborze technologii GSM 2G nie uwzględniała kierunku rozwoju technologii publicznych i ich wpływu na zakłócenie systemu kolejowego. Niestety, doświadczenia zarządców infrastruktury kolejowej innych krajów wskazują na występowaniu problemów w komunikacji, w miejscach bliskiego sąsiedztwa infrastruktury kolejowej i nadajników sieci publicznych, szczególnie szerokopasmowych (technologie 3G ÷ 5G).

W Polsce największy zarządca infrastruktury kolejowej jest w trakcie budowy sieci GSM-R. W najbliższym czasie planowane jest uruchomienie dostępu do sieci na ponad 13 tys. km linii kolejowej. Równocześnie autorowi nie są znane jakiekolwiek ustalenia, czy regulacje krajowe w zakresie rozwiązywania problemów z komunikacją w kolejowym systemie radiołączności na styku operator publiczny – operator sieci GSM-R.

W ramach prac własnych Instytutu Kolejnictwa przeprowadzono analizę otoczenia sieciowego i serię pomiarów dla wytypowanego odcinka linii kolejowej. Uzyskane wyniki zostały krótko przedstawione w niniejszym artykule. Na ich podstawie można stwierdzić, aktywność nadajników szerokopasmowych w paśmie przylegającym do pasma UIC i równocześnie nie można jednoznacznie stwierdzić, że zmierzone sygnały będą miały niekorzystny wpływ na pracę radiotelefonów kolejowych GSM-R. Tym niemniej, pierwsze odnotowane zdarzenia zakłóceń GSM-R w Polsce oraz skala wdrożenia systemu GSM-R wymagają szczególnej uwagi i działań ograniczających występowanie niekorzystnych zjawisk w obszarze kolejowym.

LITERATURA

[1] ECC Report 146. 2010. „Compatibility between GSM MCBTS and other services operating in the 900 and 1800 MHz frequency bands". CEPT.
[2] ECC Report 162. 2011. „Practical mechanism to improve the compatibility between GSM-R and public mobile networks and guidance on practical coordination". CEPT.
[3] ECC Report 229. 2015. „Guidance for improving coexistence between GSM-R and MFCN in the 900 MHz band". CEPT.
[4] EIRENE System Requirements Specification. 2016. European Integrated Railway Radio Enhanced Network. GSM-R Operators Group. UIC CODE 951. Ver. 16.0.0.
[5] EIRENE Functional Requirements Specification.2015. GSM-R Functional Group. Ver. 8.0.0
[6] Ministerstwo Infrastruktury i Budownictwa RP. 2017. „Krajowy Plan Wdrażania Technicznej Specyfikacji Interoperacyjności „Sterowanie"".
[7] O-8736-2.0/ 2014. „UIC Assessment report on GSM-R current and future radio environment". UIC.
[8] O-8740. 2013. „Report on the UIC interference field test activities in UK". UIC.
[9] Rozporządzenie 1315/2013 Parlamentu Europejskiego i Rady (UE) nr 1315/2013 z dnia 11 grudnia 2013 r. w sprawie unijnych wytycznych dotyczących rozwoju transeuropejskiej sieci transportowej i uchylające decyzję nr 661/2010/UE
[10] Rozporządzenie Komisji (UE) 2016/919 z dnia 27 maja 2016 r. w sprawie technicznej specyfikacji interoperacyjności w zakresie podsystemów „Sterowanie" systemu kolei w Unii Europejskiej (z późn. zm.)
[11] Radio Spectrum Committee: GSM-R Interferences – Contributions from delegations and ERA on issues. Statistics and best practices as a follow-up to the discussion in RSC#42. Working Document. European Commission. DG Communications Networks Content & Technology. Brussels. 26 February 2013.
[12] Sumiła M. 2016. „Risk Analysis of Interference Railway GSM-R System in Polish Conditions". Advances in Intelligent Systems and Computing. Vol. 470. pp. 469 – 479
[13] Sumiła M. 2020. „Methodical Approach to Assessing Interference in GSM-R Network". Advances in Intelligent Systems and Computing. Vol. 1032. pp. 420-429.
[14] TS 102 933-1. 2015. „Railway Telecommunications. GSM-R improved receiver parameters. Part 1: Requirements for radio reception. V2.1.1. ETSI.

METODY POMIARU PARAMETRÓW ANTEN STOSOWANYCH W KARTACH RFID

MEASUREMENT METHODS OF ANTENNA PARAMETERS USED IN RFID CARDS

Mariusz Zamłyński[1]

[1] Argo Card Sp. z o.o., Gdańsk, mzamlynski@argocard.com

DOI: 10.15199/59.2023.4.60

Streszczenie: **Artykuł przedstawia porównanie metod pomiaru parametrów anten stosowanych w kartach RFID. Weryfikację eksperymentalną oraz porównanie wyników przeprowadzono na karcie na pasmo 13,56 MHz i uzyskano spójność wyników pomiędzy metodami. Rozbieżność wyznaczonych i zmierzonych częstotliwości rezonansowych dla karty RFID nie przekracza 1,75%.**

Abstract: **The article presents a comparison of antenna parameters measurement methods used in RFID cards. Experimental verification and comparison of results was carried out on a card for the 13.56 MHz band and consistency of results between methods was obtained. The discrepancy between the determined and measured resonant frequencies for the RFID card does not exceed 1.75%.**

Słowa kluczowe: **antena, impedancja wejściowa, karta zbliżeniowa, rezonans, RFID**

Keywords: **antenna, input impedance, proximity card, resonance, RFID**

1. WSTĘP

Technologia RFID (Radio Frequency Identification) funkcjonuje w trzech pasmach częstotliwości: LF (125 kHz), HF (13,56 MHz) oraz UHF (865-868 MHz). Projektowane na nie tagi wyposażone są w anteny różniące się od siebie nie tylko pod względem wielkości, ale przede wszystkim funkcjonujące na podstawie innych zjawisk fizycznych. W przypadku anten stosowanych w kartach zbliżeniowych na pasmo 13,56 MHz rolę anteny odgrywa cewka, która zapewnia sprzężenie magnetyczne z anteną pętlową czytnika. System pracuje na częstotliwości zbliżonej do częstotliwości rezonansowej układu dwóch sprzężonych induktorów oraz podłączonej do nich równolegle pojemności chipu RFID.

Wyznaczenie parametrów obwodowych anteny RFID nie jest zagadnieniem trywialnym. W odróżnieniu od technik stosowanych w przypadku anten pracujących w systemach radiokomunikacyjnych, nie możemy tutaj w bezpośredni sposób zastosować pomiaru zespolonego współczynnika odbicia. W pierwszym podejściu antena RFID może być traktowana jako cewka, której indukcyjność można wyznaczyć poprzez pomiar jej impedancji wejściowej [4]. Po uwzględnieniu rezonansu własnego anteny RFID można wyznaczyć jej pojemność pasożytniczą, która występuje pomiędzy zwojami cewki. Wartość ta może być zmierzona przy użyciu przynajmniej dwóch technik [6]. Ostatecznie częstotliwość rezonansową karty, która zależy od indukcyjności, pojemności pasożytniczej oraz pojemności chipa, wyznacza się poprzez pomiar rezonansu układu anteny emulującej antenę czytnika RFID sprzężonej z kartą RFID [1, 3]. Przy czym podejście te nie daje informacji zwrotnej na temat wartości poszczególnych parametrów, z których rezonans ten wynika.

Przedsiębiorstwa produkujące własne anteny dla kart zbliżeniowych RFID muszą dysponować sprawdzaną metodologią wyznaczania ich parametrów na każdym etapie produkcji. Największe wyzwanie stwarza wiarygodny pomiar i kontrolowanie powtarzalności wartości pojemności pasożytniczej anteny, która zależy nie tylko od liczby zwojów, ale również grubości i rodzaju materiału z jakiego wykonana jest karta. Zadanie to jest utrudnione z względu na wrażliwość niektórych technik pomiarowych na dokładność i powtarzalność procesu kalibracji analizatora obwodów i podłączonego do niego okablowania.

W artykule zestawiono wyniki uzyskane za pomocą klasycznego pomiaru impedancji wejściowej anteny RFID z użyciem wektorowego analizatora obwodów, niepublikowanej wcześniej w tym zastosowaniu techniki pomiaru rezonansu własnego anteny RFID w układzie transmisyjnym oraz pomiaru rezonansu karty zbliżeniowej z użyciem sondy polowej.

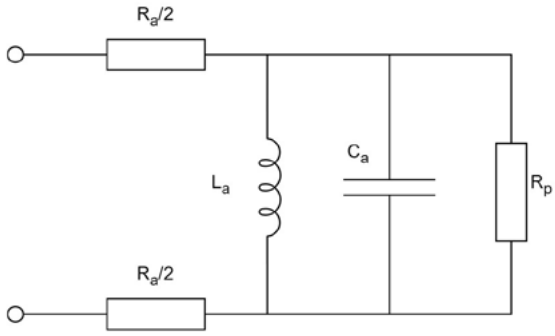
2. MODEL ANTENY RFID

Uproszczone metody projektowania anten dla kart RFID opisane przez producentów chipów RFID zakładają, że antena RFID jest idealną cewką. Model ten nie uwzględnia zjawiska występowania pojemności pasożytniczej pomiędzy zwojami anteny RFID, która razem z pojemnością chipa decyduje o częstotliwości rezonansowej karty RFID f_{rez}:

$$f_{rez} = \frac{1}{2\pi\sqrt{L_a(C_a+C_{chi})}} \quad (1)$$

gdzie L_a jest indukcyjnością anteny RFID, C_a jej pasożytniczą pojemnością, a C_{chip} pojemnością chipa RFID.

Dlatego też przy projektowaniu anteny dla karty RFID należy przyjąć model zastępczy przedstawiony na Rysunku 1, który znany jest również jako wysokoczęstotliwościowy model zastępczy cewki indukcyjnej.



Rys. 1. Model zastępczy anteny RFID.

3. METODY POMIAROWE

3.1. Pomiar impedancji wejściowej

Korzystając z modelu zastępczego anteny RFID przedstawionego na Rysunku 1 można przyjąć, że wartości poszczególnych elementów mogą być wyznaczone na podstawie znajomości przebiegu impedancji wejściowej anteny w funkcji częstotliwości. Jej pomiar może zostać zrealizowany za pomocą Wektorowego Analizatora Obwodów z obiektem badań podłączonym do jednego z portów urządzenia. W takim przypadku wartości poszczególnych elementów są wyznaczane jako:

- R_a – impedancja szeregowa anteny – uśredniona wartość części rzeczywistej impedancji wejściowej w zakresie niskich częstotliwości (np. do 500 kHz),
- L_a – indukcyjność anteny – na podstawie reaktancji zmierzonej do 1 MHz (poniżej tej częstotliwości można pominąć wpływ pasożytniczej pojemności),
- f_{SRF} – częstotliwość rezonansu własnego anteny wyznaczona na podstawie warunku zerowej wartości części urojonej impedancji wejściowej,
- R_p – rezystancja upływu – wartość modułu impedancji wejściowej na częstotliwości rezonansu własnego anteny.

Natomiast pasożytnicza pojemność wynika z wyznaczonej częstotliwości rezonansowej oraz indukcyjności anteny:

$$C_a = \frac{1}{4\pi^2 \cdot L_a \cdot f_{SFR}^2} \tag{2}$$

Praktycznie antenę RFID można podłączyć do VNA jedynie za pomocą przystosowanej do tego celu sondy. Najprostszym typem tego typu sondy jest krótki odcinek przewodu koncentrycznego z dwoma pinami. Przy czym procedura pomiarowa powinna zapewniać pomiar macierzy rozproszenia w płaszczyźnie odniesienia zdefiniowanej przez połączenie sondy z anteną RFID. W tym celu wystarczy, żeby podczas kalibracji poszczególne elementy zestawu kalibracyjnego typu SOL (z ang. Short Open Load) były przykładane do końca sondy.

Główną niedogodnością tej metody jest wpływ błędu kalibracji długości elektrycznej sondy na wartość częstotliwości rezonansowej.

3.2. Pomiar rezonansu w układzie transmisyjnym

Sama antena RFID, z względu na swoją nieidealność, jak każda inna cewka, jest równoległym układem rezonansowym LC. Przedstawione na Rysunku 2 szeregowe wpięcie takiego układu pomiędzy dwie sondy z połączonymi ekranami pozwala na precyzyjne wyznaczenie rezonansu własnego mierzonego obiektu. Na podstawie zmierzonej wartości częstotliwości rezonansowej wyznacza się pojemność C_a, a wartość modułu parametru s_{21} na częstotliwości rezonansu własnego cewki pozwala na wyznaczenie R_p:

$$R_p = 2Z_0\left(\frac{1}{|s_{21}|} - 1\right) \tag{3}$$

gdzie Z_0 jest impedancją odniesienia, dla jakiej została zmierzona macierz rozproszenia.

W metodzie tej pomiar rezonansu własnego anteny RFID zaburzony jest przez pojemność dwuportowej sondy, która występuje pomiędzy jej pinami. Wartość tej pojemności można wyznaczyć na podstawie pomiaru strat transmisji rozwartej sondy, co odpowiada pomiarowi szeregowo wpiętego kondensatora.

Związek pomiędzy transmitancją, a szeregową pojemnością wynika ze strat niedopasowania:

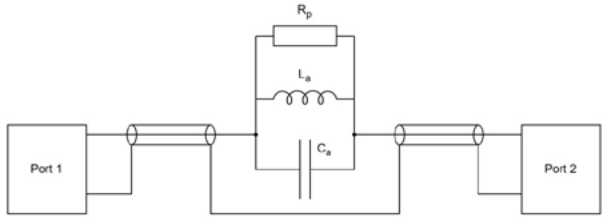
$$|s_{21}|^2 = 1 - \left|\frac{1/j\omega C}{\frac{1}{j\omega C} + 2Z_0}\right|^2 \tag{4}$$

Po przekształceniach uzyskujemy:

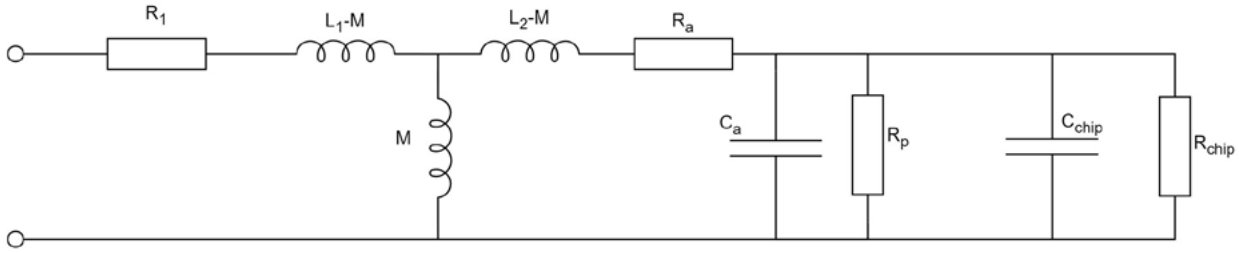
$$C_{sondy} = \frac{\sqrt{\frac{1}{1-|s_{21}|^2}-1}}{2Z_0\omega} \tag{5}$$

gdzie współczynnik $|s_{21}|$ odpowiada stratom transmisji na pulsacji ω przy pomiarze rozwartej sondy dwuportowej.

Weryfikacja poprawności kompensacji pojemności sondy pomiarowej powinna być przeprowadzona poprzez pomiar z użyciem dwóch sond o różnym rozstawie pinów. Przy poprawnie wykonanych pomiarach wynik pomiaru (tj. pojemność badanego elementu, np. chipa) powinien być w obu przypadkach zbieżny.



Rys. 2. Układ transmisyjny do pomiaru częstotliwości rezonansowej.



Rys. 3. Model zastępczy układu anteny czytnika RFID sprzężonego z kartą zbliżeniową.

3.3. Pomiar rezonansu układu zbliżeniowego

Na Rysunku 3 przedstawiono układ zastępczy anteny czytnika RFID sprzężonego z kartą zbliżeniową. Pomiar częstotliwości rezonansowej karty RFID odbywa się za pomocą następującej procedury [1, 3]:

1. Pomiar impedancji wejściowej anteny czytnika RFID bez karty zbliżeniowej,
2. Zapisanie w pamięci urządzenia przebiegu części rzeczywistej impedancji wejściowej,
3. Ustawienie trybu matematycznego odejmującego bieżące wyniki pomiarów od przebiegu zapisanego w kroku 2,
4. Przyłożenie karty RFID do anteny czytnika,
5. Odczytanie częstotliwości rezonansowej z warunku maksimum części rzeczywistej impedancji wejściowej.

4. WYNIKI EKSPERYMENTU

4.1. Induktor

Częstotliwość rezonansu własnego cewki 1025-36G zmierzona techniką pomiaru transmisyjnego wynosi 106,76 MHz. Parametry wyznaczone na podstawie pomiarów z użyciem sondy jednoportowej znajdują się w Tabeli 1. Wynika z nich, że przy założeniu indukcyjności anteny równej 4,665 uH, jej pojemność pasożytnicza zmierzona techniką dwuportową wynosi 0,48 pF. Wynik ten w znacznym stopniu odbiega od wartości wyznaczonych na podstawie rezonansu własnego zmierzonego techniką bezpośredniego pomiaru impedancji wejściowej przy użyciu dwóch metod kalibracji (bezpośredniej kalibracji oraz za pomocą funkcji ‘port extention’).

4.2. Antena RFID

Na potrzeby przeprowadzenia weryfikacji metod pomiarowych przygotowano przedstawiony na Rysunku 4 prototyp anteny RFID, który został przystosowany do pomiaru z użyciem sond kontaktowych. Antena zgodna z Klasą 1 normy ISO/IEC 14443-1 została wykonana z miedzianego drutu o średnicy 100 um oraz posiada 5 przylegających do siebie zwojów.

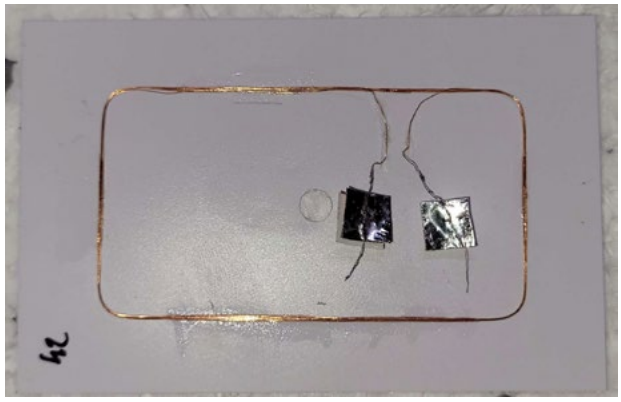
Częstotliwość rezonansu własnego anteny zmierzona techniką transmisyjną wynosi 34,15 MHz. Parametry wyznaczone na podstawie pomiarów z użyciem sondy jednoportowej znajdują się w Tabeli 2. Wynika z nich, że przy założeniu indukcyjności anteny równej 4,66 uH, jej pojemność pasożytnicza zmierzona techniką dwuportową wynosi 4,66 pF. Wynik ten w znacznym stopniu odbiega od wartości wyznaczonych na podstawie rezonansu własnego zmierzonego techniką bezpośredniego pomiaru impedancji wejściowej. Różnica sięga 17%.

Tab. 1. Parametry cewki 1025-36G

Parametr	*Kalibracja bezpośrednia sondy*	*Kalibracja z ‘port extention’*
Indukcyjność L_a	4,66 uH	4,67 uH
Rezystancja szeregowa R_a	1,03 Ohm	1,15 Ohm
Częstotliwość rezonansu własnego	61,80 MHz	89,84 MHz
Pojemność pasożytnicza C_a	1,42 pF	0,67 pF
Rezystancja upływu R_p	49,19 kOhm	40,29 kOhm

Tab. 2. Parametry anteny RFID

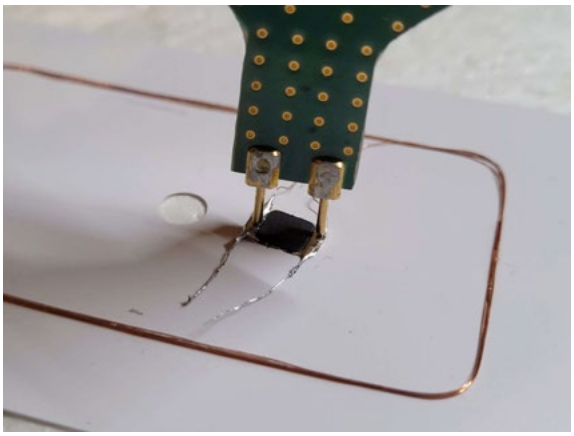
Parametr	*Kalibracja bezpośrednia sondy*	*Kalibracja z ‘port extention’*
Indukcyjność L_a	4,63 uH	4,69 uH
Rezystancja szeregowa R_a	2,18 Ohm	2,25 Ohm
Częstotliwość rezonansu własnego	30,83 MHz	31,42 MHz
Pojemność pasożytnicza C_a	5,76 pF	5,47 pF
Rezystancja upływu R_p	33,14 kOhm	30,09 kOhm



Rys. 4. Antena RFID przystosowana do pomiaru sondami kontaktowymi.

Tab. 3. Porównanie częstotliwości rezonansowych karty RFID wyznaczonych poszczególnymi metodami

Technika	*Częstotliwość rezonansowa*	*Różnica od rezonansu zmierzonego sondą zbliżeniową*
Wyliczenie na podstawie parametrów zmierzonych przez pomiar impedancji wejściowej (kalibracja bezpośrednia sondy)	16,50 MHz	-0,78%
Wyliczenie na podstawie parametrów zmierzonych przez pomiar impedancji wejściowej (kalibracja z 'port extention')	16,52 MHz	-0,66%
Wyliczenie na podstawie parametrów zmierzonych przez pomiar transmisyjny	16,92 MHz	1,74%
Pomiar w układzie transmisyjnym	16,87 MHz	1,44%
Pomiar rezonansu układu zbliżeniowego	16,63 MHz	



Rys. 5. Prototyp karty RFID podczas pomiaru rezonansu techniką transmisyjną.

4.3. Prototyp karty RFID

Kolejnym obiektem badań, na którym przeprowadzono weryfikację omówionych metod, jest prototyp karty zbliżeniowej przedstawiony na Rysunku 5. Do omówionej w poprzednim punkcie anteny RFID przylutowano chip firmy NXP o nominalnej pojemności 17 pF. Przy czym, pojemność ta występuje w warunku pobudzenia chipu sygnałem o napięciu 2V RMS. Warunek ten jest trudny do spełnienia przy pomiarach z użyciem VNA [2]. Dlatego też przed przylutowaniem chipu została zmierzona jego pojemność, która wyniosła 14,33 pF. Wartość ta jest zgodna z oczekiwaniami, ponieważ wyłączony chip RFID charakteryzuje się mniejszą pojemnością od wartości nominalnej w warunkach docelowej pracy [2]. W związku z tym oczekiwana częstotliwość rezonansowa prototypu karty wynosi 16,92 MHz.

Częstotliwość rezonansowa prototypu karty RFID zmierzona techniką sondy dwuportowej przy mocy sygnału z generatora równej -10 dBm wynosi 16,87 MHz. Oznacza to, że technika ta dostarcza spójne wyniki w zakresie częstotliwości rezonansowej karty RFID, rezonansu własnego anteny RFID oraz pojemności chipa.

Z racji tego, że pomiar częstotliwości rezonansowej metodą kontaktową przeprowadzono w warunkach wyłączenia chipu RFID, to pomiar z użyciem sondy polowej zrealizowano przy poziomie mocy generatora VNA równym -40 dBm. Dzięki czemu uniknięto uzależnienia wyników pomiaru od odległości pomiędzy mierzoną kartą RFID, a anteną sondy. Korzystając z warunku lokalnego maksimum części rzeczywistej impedancji wejściowej anteny pętlowej sprzężonej z prototypem karty RFID uzyskano wynik 16,63 MHz.

5. PODSUMOWANIE

Porównanie częstotliwości rezonansowych prototypowej karty RFID uzyskanych zarówno na podstawie parametrów anteny RFID wyznaczonych poszczególnymi metodami oraz pomiaru samego rezonansu karty przy użyciu dwóch technik przedstawiono w Tabeli 3. Jako metodę referencyjną należy wskazać technikę rezonansu układu zbliżeniowego. Po porównaniu uzyskanych wartości można wywnioskować, że wszystkie opisane metody dostarczają spójny rezultat końcowy.

Jednakże należy mieć na uwadze wrażliwość metody pomiaru impedancji wejściowej na kalibrację sondy. W przypadku anteny RFID rozbieżność w uzyskanych rezonansach własnych oraz wynikających z nich wartościach pojemności pasożytniczej okazała się mało znacząca. Jednakże w przypadku błędnie przeprowadzonej kalibracji bądź też przy pomiarze cewek o wysokich rezonansach własnych różnice te mogą przekraczać 15%.

Praca powstała w ramach projektu „Innowacyjna technologia produkcji ulepszonych kart zbliżeniowych z układami elektronicznymi i antenami na własnych podłożach" współfinansowanego z Funduszy Europejskich w ramach umowy.

LITERATURA

[1] Bitschnau Martin. 2016. „Analysis of Quality Factor and Resonance Frequency Measurements of RFID Transponders". University of Applied Sciences Vorarlberg

[2] Gebhart, Michael, Johannes Bruckbauer, Martin Gossar. 2010. „Chip impedance characterization for contactless proximity personal cards". *7th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP 2010).*

[3] Hämmerle Florian, Bitschnau Martin. 2017. „Contactless RFID Tag Measurements". *Bode 100 - Application Note*

[4] NXP. 2022, „AN12342 Card Coil Design Guide for MIFARE DESFire Light".

[5] Pesic, M, Gruber, J, Rampetzreiter, S, Witschnig, H, Arthaber, H., 2019. „A precise resonance frequency measurement method based on ISO-standardized setups for contactless chip cards". *Int J RF Microw Comput Aided Eng.*

[6] Walker Brian. 2019. „Make Accurate Impedance Measurements Using a VNA". *Microwaves & RF*

INTEGRACJA O-RAN ORAZ XPON NA WARSTWIE OPTYCZNYCH ZASOBÓW ODN

INTEGRATION OF O-RAN AND XPON ON THE OPTICAL RESOURCE LAYER OF THE ODN

Zbigniew Zakrzewski

Politechnika Bydgoska im. Jana i Jędrzeja Śniadeckich, Bydgoszcz, zbizak@pbs.edu.pl

DOI: 10.15199/59.2023.4.61

Streszczenie: **W artykule przedstawiono wyniki analizy zajętości optycznych zasobów sieci ODN przez potencjalne systemy xPON. Przedstawiono pasma optyczne możliwe do bezpośredniego wykorzystania w łączu O-RAN fronthaul. Zaprezentowano propozycje fizycznego podłączania optycznej sieci fronthaul do ODN tak, aby systemy mogły pracować niezależnie. Wykonano przykładowe obliczenia wskazujące maksymalny zasięg łącza fronthaul o wybranych szybkościach CPRI/eCPRI. Nakreślono również problematykę związaną z pozapasmowym przenikaniem sygnałów pochodzących z niedoskonałych źródeł optycznej nośnej.**

Abstract: **The paper presents the results of the analysis of ODN resource occupancy by potential xPON systems. Optical bands that can be directly used in the O-RAN fronthaul link are presented. Proposals for physically connecting the optical fronthaul network to the ODN were presented so that the systems could work independently. Exemplary calculations were made indicating the maximum range of the fronthaul link with selected CPRI/eCPRI bitrates. Problems related to out-of-band cross-talk of signals coming from imperfect optical carrier sources are also outlined.**

Słowa kluczowe: **CPRI, D-RoF, DWDM poprzez CWDM, eCPRI, fronthaul poprzez ODN, O-RAN fronthaul.**

Keywords: **CPRI, D-RoF, DWDM-over-CWDM, eCPRI, fronthaul-over-ODN, O-RAN fronthaul.**

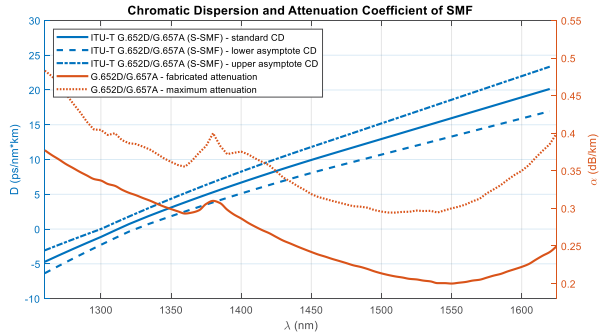
1. WSTĘP

Współczesne systemy mobilne są nacechowane wysokimi wymaganiami pod kątem przepustowości, niezawodności oraz jakości świadczonych usług. Zarówno w domenie sieci rdzeniowej, jak też w domenie sieci radiowej muszą być stosowane technologie, które są w stanie zapewnić tego rodzaju wymagania. Wszechobecność sieci dostępu radiowego, mobilnego, przenośnego lub stacjonarnego jest kluczowa przy założeniu, że technologie sieciowe i informatyczne stanowią wyznacznik rozwoju cywilizacyjnego. Aby spełnić kryterium powszechnego dostępu do usług internetowych i intranetowych, należy budować trzon szkieletowy, dystrybucyjny oraz dostępowy po stronie szerokopasmowej stacjonarnej części sieci. W ostatnich latach było realizowanych w Polsce i w Europie wiele programów rozwojowych, dzięki którym została wybudowana zasobna sieć światłowodowa o zasięgu regionalnym, podmiejskim oraz wiejskim. Utworzona infrastruktura światłowodowa zapewnia dostęp do sieci Internet nie tylko indywidualnym odbiorcom, ale także przedsiębiorstwom i parkom technologicznym.

Optyczne pasywne dystrybucyjne sieci światłowodowe ODN (Optical Distribution Network) są uznawane obecnie jako tzw. ostatnia mila w stacjonarnym dostępie do usług szerokopasmowych. Na bazie tej światłowodowej infrastruktury można uruchamiać różnorodne systemy telekomunikacyjne i teleinformatyczne, włącznie z systemami mobilnymi. Bardzo duża przepustowość oraz przezroczystość sieci światłowodowych, zwłaszcza całkowicie optycznych, umożliwia uruchamianie usług z rodziny Internetu Rzeczy globalnego (otwartego) lub prywatnego (także przemysłowego - IIoT).

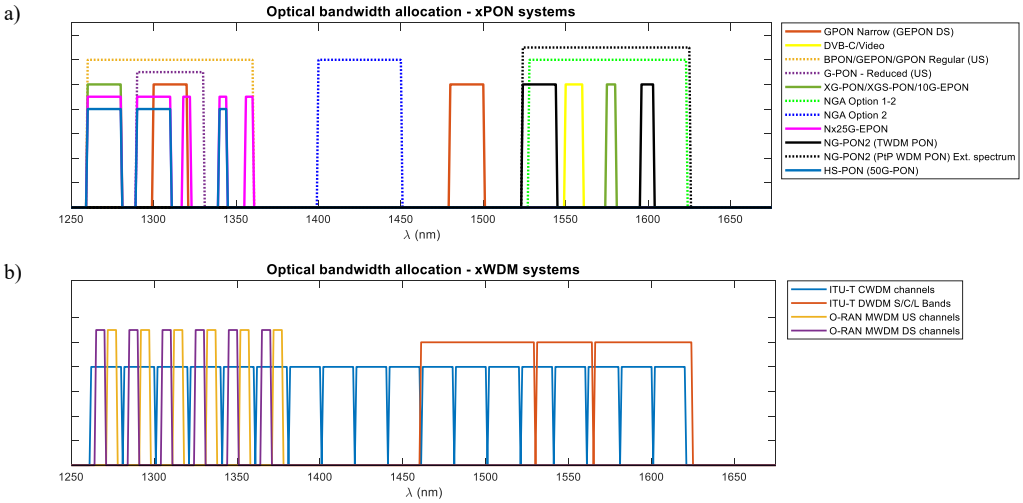
2. ANALIZA OPTYCZNYCH ZASOBÓW SIECI ODN

Pasywne sieci optyczne ODN stanowią medium transmisyjne dla optycznych systemów, które potrafią pracować w obecności pasywnych sprzęgaczy optycznych (splitters). Na przestrzeni ostatnich 30 lat systemy i sieci z rodziny xPON (Passive Optical Network) bardzo mocno się rozwinęły, ale nadal bazują na takiej samej infrastrukturze światłowodowej. Głównym medium jest światłowód jednomodowy standardu G.652 [6] lub G.657 [7], którego zakres jednomodowości jest tak samo szeroki i zawiera się w pasmie falowym 1260-1675 nm.



Rys. 1. Charakterystyka współczynnika dyspersji chromatycznej oraz tłumienności jednostkowej włókien światłowodowych standardów G.652D [5,6] oraz G.657A [7]

Na rysunku 1 przedstawiono graniczne oraz wypadkowe charakterystyki współczynników tłumienności oraz dyspersji chromatycznej włókien jednomodowych stosowanych w sieciach pasywnych. Pasma pracy systemów xPON są dostosowane do pasma jednomodowości tych światłowodów. Dyspersja chromatyczna oraz tłumienie to zasadnicze parametry wpływające na ograniczenie zasięgu łącza światłowodowego pracującego bez wzmacniania i kompensacji. Na rysunku 2 przedstawiono zbiorcze zestawienie zajętości pasm optycznych przez tzw. pasywne i aktywne systemy i sieci optyczne.



Rys. 2. Zestawienie zajętości (obsługi) dostępnych pasm optycznych przez: a) pasywne sieci optyczne [4,10,12, 13,14,16,17], b) systemy i sieci bazujące na technice zwielokrotnienia falowego WDM [8,9,19]

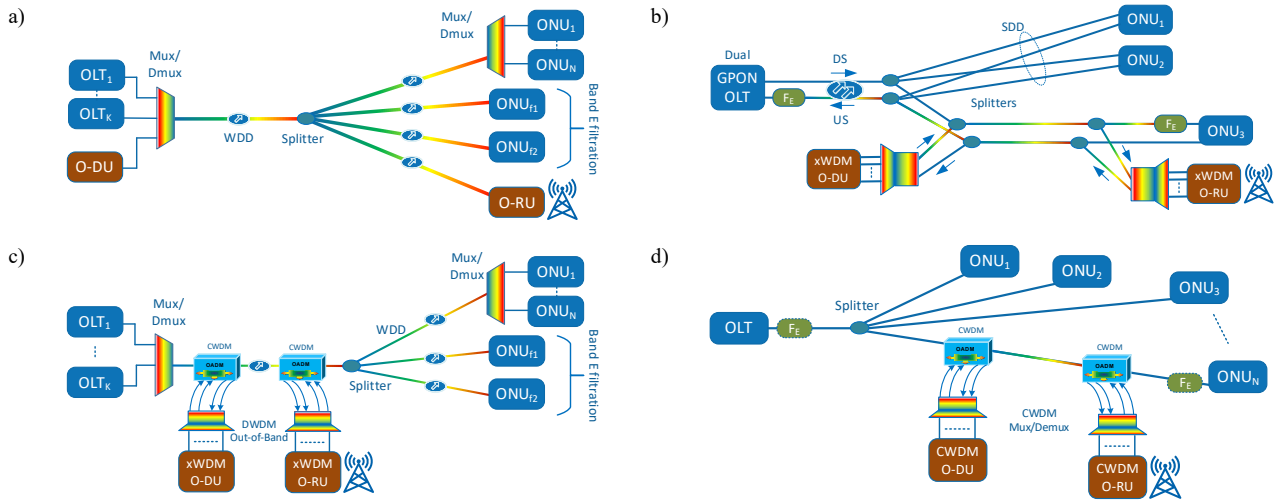
Oczywistym jest, że nie wszystkie systemy są uruchamiane naraz w tej samej sieci ODN, zatem w każdym przypadku podejście do planowania w wykorzystaniu zasobów powinno być inne. Obecnie najczęściej są uruchamiane systemy GPON [14] lub GEPON (802.3ah lub 802.3bk) [4], które zajmują takie same lub podobne zasoby (US: 1260-1360nm, DS: 1480-1500nm – przy jednym włóknie oraz US/DS: 1260-1360nm – przy dwóch włóknach). System GPON jest następcą BPON [13], w którym mogły mieć zastosowanie lasery MLM o niskiej stabilności oraz o szerokim spektrum, na co należy zwracać szczególną uwagę przy efektywnym zarządzaniu pasmem sieci ODN. Nienaruszalne pasmo 1550-1560nm jest na stałe zarezerwowane dla sieci kablowych DVB-C lub innych sygnałów video w kierunku DS. Sieci z interfejsami 10Gb/s pracują wyłącznie na jednym włóknie (DS/US) i zajmują inne zasoby, co gwarantuje koegzystencję z optycznymi sygnałami GPON/GEPON. W przypadku XG-PON [16], XGS-PON [12] lub 10G-EPON [4] jest to zakres 1260-1280nm (US) oraz 1575-1580nm (DS). Systemy z rodziny NG-PON2 [17] zostały zaprojektowane do pracy w okolicy optycznych pasm C oraz L. System TWDM PON zajmuje pasmo 1596-1603nm (DS) oraz 1524-1544nm (US) w 3 wariantach pasmowych. Dla systemu PtP WDM PON zarezerwowano pasma 1524-1625 (DS) oraz 1603-1625nm (US) w zakresach niezajętych przez inne systemy – czarne linie na rysunku 2a. Do ostatniej grupy zaliczamy systemy z interfejsami 25 i 50 Gb/s. Ze względu na minimalny współczynnik dyspersji chromatycznej we włóknach G.652/G.657 w okolicy pasma O, to właśnie w tej okolicy zarezerwowano pasma dla tych systemów. Zakresy zostały tak dobrane, aby była możliwość współpracy z GPON. HS-PON (50G-PON) [10] pracuje w pasmach 1340-1344nm (DS) oraz w trzech wariantach US zależnych od szybkości interfejsu (1260-1280nm – opcja 1 dla 12,5Gb/s, 1290-1310nm – opcja 2 (wideband) dla 12,5Gb/s oraz 1298-1302nm – opcja 2 (narrowband) dla 25Gb/s). Konkurencyjny system IEEE Nx25G-EPON (802.3ca) [4] także pracuje w paśmie O oraz w większości interfejsów gwarantuje zgodność z systemami GPON/GEPON/XG-PON/10G-EPON.

W regulacji [15] zdefiniowano dwa pasma przewidziane do pracy systemów następnych generacji, tj. NGA Option 1-2 oraz NGA Option 2 – rysunek 2a). Pierwsze z nich już jest zarezerwowane przez systemy podstawowe oraz o zwiększonej szybkości. Pasmo Option 2 (większość optycznego pasma E) jest nadal wolne oraz od strony fal krótkich i długich jest otoczone szerokimi odstępami ochronnymi, których należy przestrzegać tylko w przypadku, gdy mają zastosowanie mało stabilne lasery interfejsów GPON/GEPON. Wobec tego warto jest zainteresować się tym nadal wolnym zakresem.

W drugiej grupie systemów światłowodowych mamy rozwiązania oparte wyłącznie na technice WDM. Technika WDM w wąskim zakresie jest także stosowana w interfejsach xPON dla liniowych strumieni 10, 12,5, 25 oraz 50 Gb/s. Na rysunku 2b) widać, że system DWDM [8] pracuje w optycznych pasmach S, C oraz L, podczas gdy system CWDM [9] zajmuje praktycznie cały zakres jednomodowości światłowodu standardu G.652/G.657. Do zastosowań w optycznych sieciach Xhaul zaproponowano rozwiązanie MWDM [19]. Sieć ta jest dedykowana do transportowania strumieni CPRI [2] oraz eCPRI [3] w sieci O-RAN na poziomie interfejsów Fronthaul (FH) oraz Midhaul (MH). Wprowadzone 6 par kanałów daje możliwość dopasowania zakresów falowych tak, aby efektywnie zintegrować strumienie Xhaul ze strumieniami GPON/XG-PON. Jednak potrzeba zachowania odstępów ochronnych wynikających z niskiej stabilności laserów, może wykluczyć koegzystencję.

3. FIZYCZNA INTEGRACJA SIECI O-RAN ORAZ ODN

Sprzęganie sieci zewnętrznych z systemami xPON może odbywać się w lokalizacji aktywnych urządzeń węzłowych OLT/ONU (rys. 3a) lub w obszarze sieci ODN. W przypadku potrzeby podłączenia łącza O-RAN do sieci światłowodowej poza punktami OLT/ONU, należy wziąć pod uwagę nie tylko zajętość optycznych zasobów, ale także potrzebę wpięcia pasywnych układów, które wniosą dodatkowe tłumienie wszystkich sygnałów.



Rys. 3. Schematy podłączania sieci O-RAN do ODN z uwzględnieniem pracy systemów xPON (wykorzystanie pasma E): a) integracja w puntach OLT/ONU [11], b) wykorzystanie zasobów w dystrybucyjnej części dualnej sieci ODN, c) transport DWDM eCPRI poprzez CWDM w fiderze optycznym, d) transport CWDM CPRI w dystrybucyjnej części sieci.

W przypadku splitterów, będzie to wzrost tłumienia o około 4 dB (rys. 3b). Obecność pasywnych OADMów (rys. 3c oraz 3d) spowoduje wzrost tłumienia przelotowego nawet o dodatkowe 5-7 dB. Na rysunku 3 (b, c, d) przedstawiono trzy przykładowe propozycje podłączenia optycznego łącza O-RAN do sieci ODN, z uwzględnieniem konieczności zwiększenia klasy tłumieniowej łączy xPON o jeden stopień, tj. na przykład przejście z klasy B+ na C+ lub z klasy C+ na D (w odniesieniu do klas GPON [14]). Przy tym należy zwrócić uwagę, że optyczne pasmo E [18] leży w zakresie absorpcji jonów OH^-. W przypadku włókien z rodziny OS2 tłumienność ta jest znacznie ograniczona, ale nadal może osiągać wartość dochodzącą do 0,4 dB/km – rys.1.

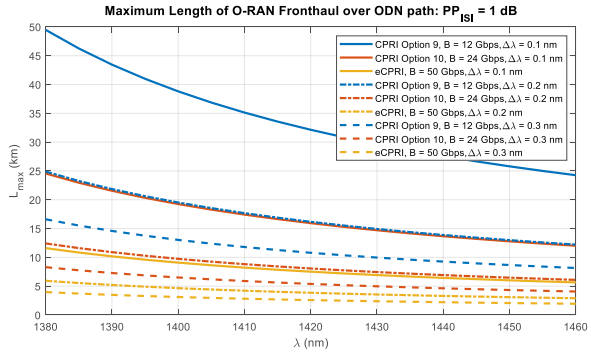
Kolejnym zagadnieniem jest sprawdzenie dopuszczalnego zasięgu łącza fronthaul sieci O-RAN, które będzie ograniczone dyspersją chromatyczną występującą w optycznym paśmie E w światłowodzie stosowanym w sieci ODN, czyli G.652D/G.657A. W podstawowym interfejsie optycznym ma zastosowanie format NRZ-OOK, dla którego współczynnik wypełnienia impulsu można obliczyć na podstawie dopuszczalnego zmniejszenia czułości odbiornika dla określonej interferencji międzysymbolowej PP_{ISI} [18]:

$$\varepsilon = \sqrt{\left(10^{\frac{PP_{ISI}}{5}} - 1\right)} \Big/ 2\pi \qquad (1)$$

Dopuszczalny zasięg łącza o określonej przepływności bitowej możemy wyznaczyć przy pomocy zależności [1,18]:

$$L_{\max} = \frac{\varepsilon c}{D\lambda^2 B\sqrt{\left(\frac{B}{\pi f}\right)^2 + \sigma_v^2}}, \qquad (2)$$

gdzie odpowiednio D – współczynnik dyspersji chromatycznej, λ - długość fali optycznej nośnej, B – bitowa przepływność łącza, f – wypełnienie szczeliny czasowej, σ_v – 20-dB częstotliwościowa szerokość widma lasera. Na podstawie zależności (1) i (2) wyznaczymy maksymalny zasięg łącza fronthaul D-RoF. Wyniki tych obliczeń zostały przedstawione na rysunku 4.



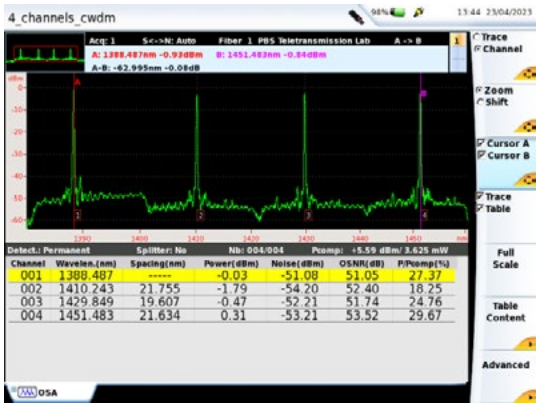
Rys. 4. Maksymalny zasięg łącza fronthaul D-RoF pracującego w optycznym paśmie E w formacie NRZ-OOK.

Wyniki obliczeń pokazują, że kluczowe jest stosowanie laserów o wąskiej linii widmowej. W przypadku poza systemowego łącza DWDM poprzez CWDM (rys. 3c) nie jest to problemem, jednak przy zastosowaniu tanich laserów CWDM (rys. 3d) szerokość spektralna jest istotna.

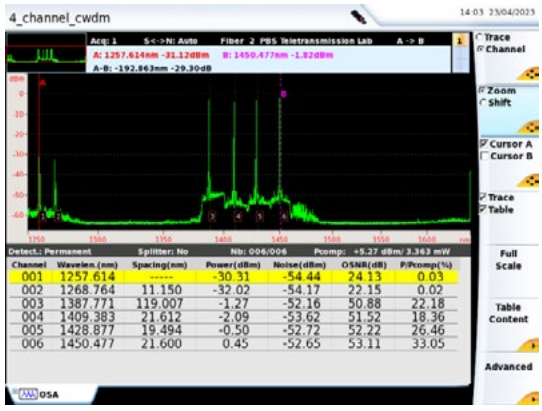
W celu pewnego sprzężenia łącza fronthaul z siecią ODN, w której pracuje jeden lub więcej systemów xPON, należy pamiętać o systemowych częstotliwościowych odstępach ochronnych, które mogą pokrywać się z zakresem pracy filtrów optycznych wbudowanych w OADMy. W przypadku uruchomienia w optycznym paśmie E czterech kanałów CWDM (rys. 5), zarówno od strony fal krótkich, jak i długich, część zakresów ochronnych (rys. 2) będzie pokrywała się z zakresami pracy filtrów kanałów o numerach 7 i 10 (rys. 2b) – na rysunku 5 przedstawiającym zrzut z OSA są to kanały 1 i 4.

Innym zagadnieniem jest uwzględnienie pozapasmowej emisji laserów stosowanych w systemie CWDM. Zważywszy, że przy koegzystencji O-RAN fronthaul z xPON wykorzystujemy pełny zakres jednomodowości

sieci ODN, musimy zadbać, aby niepożądane sygnały nie interferowały między sobą w stopniu zakłócającym ich pracę.



Rys. 5. Wyniki spektralnych pomiarów 4-kanałowego sygnału CWDM, wykonanych w optycznym paśmie E w lokalizacji za multiplekserem przelotowym OADM.



Rys. 6. Wyniki spektralnych pomiarów wykonanych za OADM w pełnym zakresie jednomodowości toru światłowodowego, przy włączonych 4 sygnałach optycznych z siatki CWDM w optycznym paśmie E.

Na rysunku 6 przedstawiono przykład emisji pozapasmowej (oznaczone jako kanały o numerach 1 i 2) pochodzącej od sygnałów CWDM, która nieco przekracza dopuszczalną wartość SMSR wynoszącą 30 dB [14]. Ta niepożądana emisja nie została wyeliminowana przez filtry multipleksera OADM.

4. WNIOSKI I DALSZE BADANIA

Wyniki przeprowadzonej analizy pokazują, że występują jeszcze wolne zakresy w optycznym paśmie sieci ODN, które można wykorzystać do niezależnej transmisji w przypadku przelotowego wpięcia łącza O-RAN fronthaul. Jest to na przykład pasmo E, w którym mieszczą się 4 kanały CWDM. W przypadku potrzeby wprowadzenia większej liczby strumieni do O-RU pracującego w formacie AAU massive-MIMO, można zastosować technikę DWDM-over-CWDM (propozycja na rysunku 3c), ale przy uruchomieniu kanałów DWDM nie leżących w zakresie pracy systemu. Bardzo ważnym aspektem dobrze zaprojektowanego łącza fronthaul, bazującego na pracującej sieci ODN, jest zwrócenie uwagi na niepożądane sygnały pozapasmowe, zwłaszcza te leżące poza zakresem filtrów stosowanych na całej drodze OLT-ONU. Jest to zagadnienie, któremu będą poświęcone kolejne badania oraz symulacje komputerowe. Na rysunkach 4 i 5 przedstawiono zrzuty z optycznego analizatora widma, które prezentują spektrum tworzone przez 4 strumienie CWDM o szybkości 1Gb/s każdy. Jak widać kanały CWDM są na tyle szerokie, że można w nich umieścić kilka szybszych strumieni, ale przy wykorzystaniu dużo stabilniejszych laserów SLM. Jednocześnie na spektrogramie z rysunku 6 widać wspomniane wcześniej niepożądane sygnały dla fal krótkich, które mogą np. zakłócić pracę strumieni GPON z kierunku US. W przypadku sygnałów A-RoF, należy wykonać obliczenia i symulacje w zakresie DIPP, zwłaszcza w przypadku formatu DSB-RFoF [20].

LITERATURA

[1] Agrawal, G.P. 1997. *Fiber-Optic Communication Systems*, 2nd Edition, New York, John Wiley & Sons, Inc.
[2] CPRI Industry Forum (Ericsson, Huawei, NEC, and Nokia). 2015. CPRI Specification ver. 7.0.
[3] CPRI Industry Forum (Ericsson, Huawei, NEC, and Nokia). 2019. eCPRI Specification ver. 2.0.
[4] IEEE 802.3. 2022. IEEE Standard for Ethernet.
[5] ITU-T, G.650.1. 2020. Definitions and test methods for linear, deterministic attributes of single-mode fibre and cable.
[6] ITU-T, G.652. 2016. Characteristics of a single-mode optical fibre and cable.
[7] ITU-T, G.657. 2016. Characteristics of a bending-loss insensitive single-mode optical fibre and cable.
[8] ITU-T, G.694.1. 2020. Spectral grids for WDM applications: DWDM frequency grid.
[9] ITU-T, G.694.2. 2003. Spectral grids for WDM applications: CWDM wavelength grid.
[10] ITU-T, G.9804.3, 2021. 50-Gigabit-capable passive optical networks (50G-PON): Physical media dependent (PMD) layer specification.
[11] ITU-T, G.9805. 2022. Coexistence of passive optical network systems.
[12] ITU-T, G.9807.1, Cor.1. 2020. 10-Gigabit-capable symmetric passive optical network (XGS-PON).
[13] ITU-T, G.983.1. 2005. Broadband optical access systems based on Passive Optical Networks (PON).
[14] ITU-T, G.984.2. 2019. Gigabit-capable passive optical networks (GPON): Physical media dependent (PMD) layer specification.
[15] ITU-T, G.984.5. 2022. Gigabit-capable passive optical networks (G-PON): Enhancement band.
[16] ITU-T, G.987.2, Amd.2. 2020. 10-Gigabit-capable passive optical networks (XG-PON): Physical media dependent (PMD) layer specification.
[17] ITU-T, G.989.2, Amd.1. 2016. 40-Gigabit-capable passive optical networks 2 (NG-PON2): Physical media dependent (PMD) layer specification.
[18] ITU-T, Supp. 39 to G-series recomm. Optical system design and engineering considerations.
[19] O-RAN Alliance. 2023. O-RAN Open X-haul Transport WG9. WDM-based Fronthaul Transport 3.0. Technical Specification ORAN-WG9.
[20] Zakrzewski, Zbigniew. 2021. "Optical Channel Selection Avoiding DIPP in DSB-RFoF Fronthaul Interface". *Entropy* 23(11): 1554.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

WYKORZYSTANIE REKONFIGUROWALNYCH MATRYC ANTENOWYCH WRAZ Z INFORMACJĄ KONTEKSTOWĄ

UTILIZATION OF RECONFIGURABLE INTELLIGENT SURFACES WITH CONTEXT INFORMATION

Łukasz Kułacz[1], Adrian Kliks[2]

[1] Politechnika Poznańska, Poznań, lukasz.kulacz@put.poznan.pl
[2] Politechnika Poznańska, Poznań, adrian.kliks@put.poznan.pl

DOI: 10.15199/59.2023.4.62

Streszczenie: Rekonfigurowale matryce antenowe mogą zostać użyte z powodzeniem do sterowania środowiskiem radiowym. Proste sterowanie kątem odbicia sygnału od powierzchni pozwala na maksymalizację lub minimalizację odbieranej mocy w określonych miejscach. W artykule przedstawiono symulacje, gdzie możliwy jest odbiór sygnału w miejscu, w którym nie było to możliwe, wykrycie zajętości widma w miejscu, w którym detektor nie był w stanie dokonać prawidłowej detekcji, czy zminimalizowanie zakłóceń w konkretnym odbiorniku.

Abstract: Reconfigurable intelligent surfaces can be successfully used to control the radio environment. Simple control of the reflection angle of the signal from the surface allows maximization or minimization of the received power in specific places. The paper presents simulations where it is possible to receive a signal in a place where it was not possible, to detect the occupancy of the spectrum in a place where the sensor was unable to make correct detection or to minimize interference in a specific receiver.

Słowa kluczowe: detekcja zajętości widma, informacja kontekstowa, propagacja sygnału, rekonfigurowalne matryce antenowe

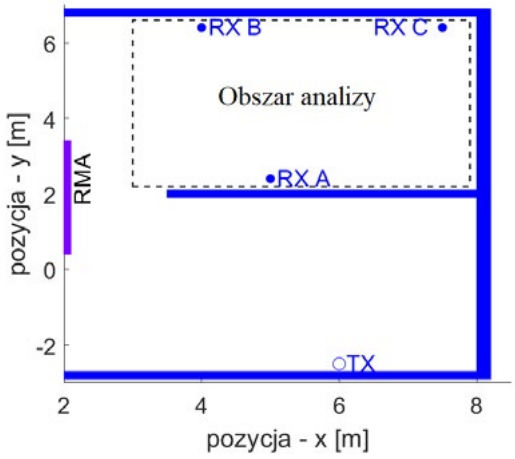
Keywords: context information, reconfigurable intelligent surfaces, signal propagation, spectrum occupancy detection.

1. WSTĘP

Rozwój komunikacji bezprzewodowej w ostatnich czasach opiera się w bardzo często na wprowadzaniu do sieci coraz większego poziomu inteligencji. Pod pojęciem wspomnianej inteligencji uznaje się między innymi systemy dynamicznego dostępu do widma, które często wymagają dostępu do dużej ilości informacji kontekstowej o środowisku i użytkownikach sieci [1]. Informacje te są niezbędne do podejmowania optymalnych decyzji – w oparciu o wiele parametrów (niekoniecznie związanych bezpośrednio z samą siecią) oraz ich historyczne wartości. Dane historyczne pozwalają między innymi wychwycić trendy w zachowaniach użytkowników sieci, a tym samym umożliwiają przewidywanie nadchodzących wydarzeń. Środowisko miejskie jest niezwykle wymagające dla projektantów systemów bezprzewodowych, gdzie przesyłany sygnał poddawany jest wielu zjawiskom propagacyjnym, tj. wielokrotnym odbiciom czy rozproszeniom. Ogromna liczba i różnorodność budynków i obiektów wpływających łącznie na propagację sygnału tworzy bardzo skomplikowany kanał radiowy. Dodatkowo w środowisku miejskim występuje ogromna liczba użytkowników na stosunkowo niewielkim obszarze. Połączenie tych dwóch faktów stanowi spore wyzwanie w kontekście działania systemów bezprzewodowych. Warto zauważyć, że generalnie odbicia sygnału są nieuniknione i choć utrudniają analizę propagacji sygnału, to można uznać je za zjawisko korzystne - ponieważ sygnał ma możliwość dotarcia do odbiornika, który nie znajduje się bezpośrednio w polu widzenia nadajnika. W tym kontekście szczególnie ciekawym podejściem jest popularny i rozważany w ostatnim czasie temat Rekonfigurowalnych Matryc Antenowych, RMA (ang. Reconfigurable Intelligent Surfaces, RIS) [3]. RMA pozwalają wpływać na proces odbicia fal radiowych na swojej powierzchni. Oznacza to, że możliwa jest modyfikacja podstawowej zasady, w której kąt padania fali jest równy kątowi odbicia fali. Stwarza to ogromne możliwości w kreowaniu i sterowaniu środowiskiem propagacyjnym. Z jednej strony kierowanie sygnału radiowego w miejsca wcześniej niedostępne (ze względu na geometrię obiektów). Z drugiej strony usunięcie sygnału radiowego w określonym miejscu (minimalizacja zakłóceń) poprzez skierowanie go w zupełnie innym kierunku [4]. Tylko te dwie koncepcje wydają się mieć duży potencjał do wykorzystania w sieciach inteligentnych i mogą służyć do wykrywania sygnału lub poprawy jakości odbieranego sygnału (poprzez zwiększenie odbieranej mocy, jeśli jest to pożądany sygnał lub zmniejszenie odbieranej mocy, jeśli jest to niepożądany sygnał). Wykrywanie sygnału mogłoby być możliwe dla detektora, który jest w jakiś sposób ukryty przed nadajnikiem nie tylko w sensie bezpośredniej widoczności, ale także sygnałów odbitych o odpowiednio dużej mocy niezbędnej do działania algorytmów detekcji sygnału. Wskazane rozwiązania wymagają jednak opracowania nowych algorytmów, których zadaniem jest sterowanie opisaną RMA [2].

2. OPIS ANALIZOWANEGO SCENARIUSZA

W tym artykule rozważane jest wydłużone pomieszczenie, w którym między nadajnikiem sygnału a trzema odbiornikami znajduje się ściana uniemożliwiająca bezpośrednią widoczność urządzeń. W głębi pomieszczenia znajduje się jednak powierzchnia, która będzie symulować działanie RMA.



Rysunek 1: Obszar analizy

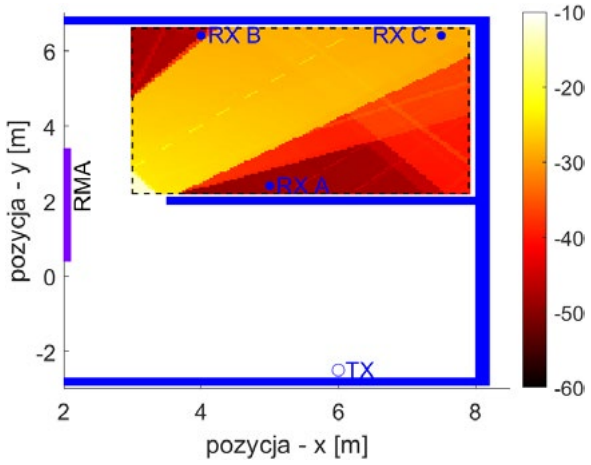
Wszystkie urządzenia znajdują się na tej samej wysokości, a wpływ sufitu i podłogi został pominięty – upraszczając analizowany przypadek do przestrzeni dwuwymiarowej. Rzut analizowanego obszaru przedstawiono na Rys. 1. Opisywane pomieszczenie zostało wybrane celowo jako demonstracja potencjalnych możliwości wykorzystania RMA. Decyzja ta podyktowana jest chęcią naświetlenia koncepcji wykorzystania RMA, a same pomysły mogą być wykorzystane w bardziej złożonych analizach, gdzie jednak znacznie trudniej byłoby wyizolować wpływ RMA (dla celu demonstracji). W artykule porównano trzy scenariusze symulacji:

- *Scenariusz 1* — zakłada brak użycia RMA, dlatego w jego miejscu umieszczony został fragment ściany. Ten scenariusz stanowi odniesienie dla pozostałych scenariuszy.
- *Scenariusz 2* — zakłada uproszczone działanie RMA, które jest niezależne od czynników zewnętrznych, tj. okresowo i naprzemiennie zmieniany jest kąt odbitego sygnału. Symuluje to użycie RMA jako płaszczyzny odbijającej (analogicznie do lustra, tylko zakresie fal elektromagnetycznych). Scenariusz ten został zaproponowany w celu wykazania potencjalnych korzyści płynących z zastosowania nawet tak prostego mechanizmu sterowania.
- *Scenariusz 3* — to sytuacja, w której użytkownicy sieci mają dostęp do informacji o bieżącym ustawieniu RMA i mają pośredni wpływ na to ustawienie. Ostatni scenariusz to przykład wykorzystania informacji kontekstowej (w tym przypadku aktualnego ustawienia RMA) do osiągnięcia bieżącego celu (np. minimalizacji zakłóceń).

W artykule przeanalizowano trzy przypadki użycia RMA sprawdzając wszystkie trzy scenariusze (w każdym przypadku użycia). W szczególności rozważono poprawę skuteczności wykrywania zajętości widma przez detektor; zwiększenie mocy sygnału odbieranego w odbiorniku; redukcja zakłóceń w odbiorniku. Podczas symulacji symulowano działanie RMA jako zmianę kąta obrotu, pod jakim zaznaczona na Rys. 1 powierzchnia (jako RMA) znajduje się względem równoległej do niej ściany (po prawiej stronie rysunku). Rozważono wartości kąta od -20 do 20 stopni z krokiem 5 stopni.

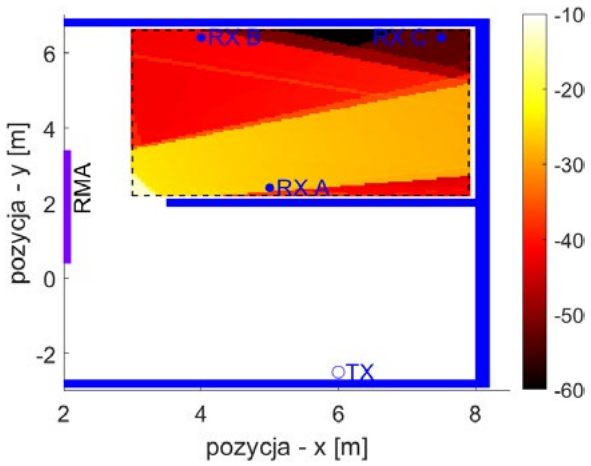
3. WYNIKI SYMULACJI

Na Rys. 2 przedstawiono rozkład przestrzenny mocy odbieranej na analizowanym obszarze dla zadanych parametrów transmisyjnych i przy wykorzystaniu modelu propagacji sygnału opartego na śledzeniu promieni. Do przeprowadzenia symulacji wykorzystane zostało oprogramowanie MATLAB (z zestawem narzędzi "Antenna Toolbox").



Rysunek 2: Odbierana moc sygnału bez RMA [dBm] (scenariusz 1)

Ze wstępnej analizy przedstawionego rozkładu wynika, że część obszaru charakteryzuje się mniejszym tłumieniem sygnału (barwy cieplejsze, odbiornik C), natomiast pozostała część obszaru ma większe tłumienie sygnału (kolory ciemniejsze, odbiorniki A i B – pozostające "w cieniu" sygnału). Moc odbierana w tej sytuacji wynika bezpośrednio z prostopadłego ułożenia ściany oznaczonej na Rys. 1 jako "RMA" – tu jednak pełniącej rolę fragmentu ściany. Można zauważyć, że odbiorniki A i B miałyby problem zarówno z odbiorem, jak i wykryciem obserwowanego sygnału, byłyby jednak pozbawione nadmiernych zakłóceń. Sytuacja jest dokładnie odwrotna dla odbiornika C. Należy jednak pamiętać, że w odwrotnej sytuacji, gdy odbiorniki A i B uznają analizowany sygnał za niepożądany, a odbiornik C za sygnał pożądany, to oddziaływanie na propagację nie byłoby korzystne.

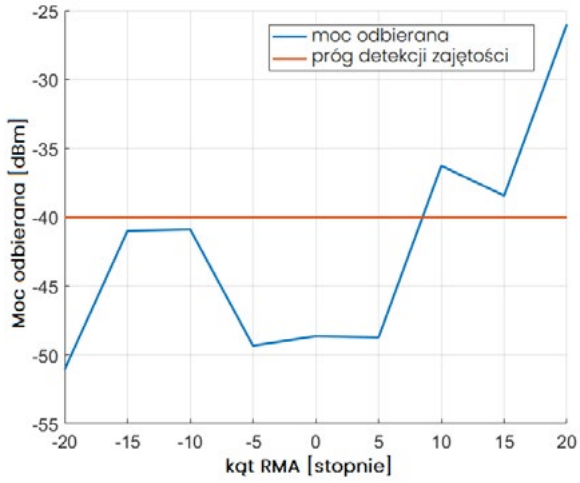


Rysunek 3: Odbierana moc sygnału z obecnym RMA [dBm] (scenariusz 3)

Dla porównania na Rys. 3 przedstawiony został rozkład przestrzenny mocy odbieranej dla sytuacji, gdy RMA jest w użyciu i obija sygnał w taki sposób, że odpowiada to obróceniu fragmentu ściany o 20 stopni. Analiza tego rozkładu ukazuje, że odbiornik A, obserwuje w tej sytuacji o wiele większą moc odbieraną w stosunku do poprzedniej analizy (przedstawionej na Rys. 1). W przeciwnej sytuacji znajduje się odbiornik B. W następnych podrozdziałach przeanalizowane zostaną kolejne przypadki użycia RMA w symulowanym środowisku.

3.1 Pierwszy przypadek użycia: wykrywanie zajętości widma

Dla pierwszego przypadku użycia tj. poprawy skuteczności detekcji zajętości widma przez odbiornik A, symulacje wykazały, że w pierwszym scenariuszu poprawna detekcja jest bardzo utrudniona (wręcz niemożliwa), w drugim scenariuszu poprawna detekcja jest możliwa przez około 31,25% czasu symulacji, a w trzecim scenariuszu, przez około 79,17% czasu istnieje możliwość poprawnej detekcji. Moc obserwowana w odbiorniku A w drugim scenariuszu (w zależności od ustawienia kąta RMA) przedstawiono na Rys. 4.

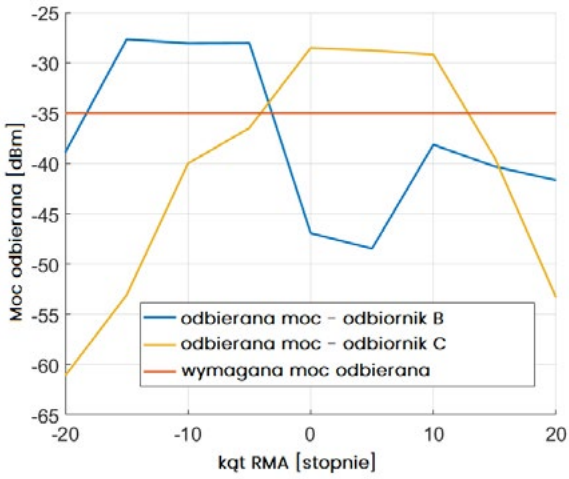


Rysunek 4: Odbierana moc w odbiorniku A (scenariusz 2)

Oczywiście ostatni scenariusz jest silnie uzależniony od zastosowanego algorytmu (i jego przeznaczenia). Jednakże nawet informacja o aktualnym ustawieniu RMA jest wystarczająca, aby detektor aktualizował decyzje o zajętości widma tylko dla odpowiednich ustawień RMA. To podejście nie jest idealne, a jego skuteczność zależy w dużej mierze od charakterystyki transmisji (częstotliwość, długość transmisji itp.). Rozwiązanie zakładające możliwość sterowania RMA nawet w bardzo prostym scenariuszu umożliwia np. okresowe "przeczesywanie" wszystkich dostępnych ustawień, a następnie utrzymywanie jednego ustawienia (lub kilku na przemian) przez określony czas. W rozważanym systemie z jednym detektorem po krótkim czasie obserwacji możliwe jest ustawienie RMA na cały pozostały czas symulacji, co pozwala osiągnąć możliwość prawidłowej detekcji przez prawie 100% czasu.

3.2 Drugi przypadek użycia: Poprawa poziomu odbieranego sygnału

Dla drugiego przypadku użycia, tj. poprawy poziomu odbieranego sygnału dla odbiorników B i C, wyniki symulacji pokazują, że w pierwszym scenariuszu odbiornik B nie jest w stanie poprawnie odebrać sygnału, podczas gdy odbiornik C może odbierać sygnał poprawnie przez cały czas. W drugim scenariuszu, gdzie ustawienie RMA jest zmieniane okresowo, przez 35,29% czasu odbiornik B odbiera sygnał o wystarczającej mocy, podczas gdy odbiornik C przez 41,18% czasu. Moc odbierana w zależności od ustawionego kąta RMA dla scenariusza drugiego została przedstawiona na Rys. 5. W tym przypadku widzimy zysk w postaci możliwości obsługi odbiornika, który znajdował się w "cieniu" sygnału i nie był w stanie odebrać sygnału bez zmiany swojego położenia. W trzecim scenariuszu przyjęto, że przez początkowy czas RMA sprawdza wszystkie dostępne ustawienia, a następnie na podstawie zgłoszeń od odbiorników naprzemiennie korzysta ze zgłoszonych (najlepszych dla poszczególnych odbiorników) ustawień. W rezultacie odbiornik B był w stanie odbierać sygnał przez 45,1% czasu, a odbiornik C przez 47,06% czasu. W tym miejscu warto wspomnieć o możliwości uwzględnienia algorytmów sterowania ruchem w celu dostosowania czasu trwania ustawień dla poszczególnych odbiorników.



Rysunek 5: Odbierana moc w odbiornikach B i C

3.3 Trzeci przypadek użycia: redukcja zakłóceń

Dla trzeciego przypadku użycia tj. redukcji zakłóceń w odbiorniku C, założono dopuszczalny próg mocy sygnału zakłócającego, powyżej którego odbiornik jest nadmiernie zakłócany. Wyniki symulacji pokazują, że w pierwszym scenariuszu przez 100% czasu odbiornik obserwuje znaczne zakłócenia (powyżej założonego progu). W drugim scenariuszu zakłócenia przekraczają próg przez 35,29% czasu. Warto w tym miejscu wspomnieć, że średni obserwowany poziom zakłóceń spada o około 3,715 dB, mediana o około 7,972 dB, a 10 percentyl o około 24,762 dB. Oczywiście 90-ty percentyl pozostaje taki sam. Znajomość ustawienia RMA mogłaby przynajmniej dostarczyć informacji o momencie wystąpienia potencjalnych zakłóceń – co można np. wykorzystać w procesie planowania transmisji. W trzecim scenariuszu ponownie zakładamy określony

czas na sprawdzenie wszystkich ustawień RMA, a następnie na podstawie raportu odbiornika C ustawienie konfiguracji powodującej minimalny poziom zakłóceń. Dzięki temu udało się skrócić czas zakłóceń powyżej założonego progu z 100% czasu do 19,61% czasu. Wyznaczony czas ponownie jest mocno zależny od danej konfiguracji i obecności innych urządzeń.

Należy pamiętać, że przeprowadzone symulacje zakładały stały w czasie schemat transmisji podczas całego eksperymentu. W przypadku jedynie chwilowych transmisji sygnału dla wszystkich trzech przypadków użycia byłoby to dodatkowe wyzwanie dla algorytmu sterującego pracą RMA.

3.4 Wszystkie przypadki użycia łącznie

Dodatkowo przeprowadzono eksperyment symulacyjny łączący wszystkie trzy przedstawione przypadki użycia, tj. odbiornik A jest detektorem wykrywającym zajętość widma, odbiornik B jest odbiornikiem analizowanego sygnału (sygnału pożądanego), a odbiornik C nie jest odbiornikiem analizowanego sygnału (sygnał niepożądany). Zauważmy, że w scenariuszu referencyjnym cele żadnego z odbiorników nie są spełnione. Następnie zastosowano RMA sterowany według następującego schematu: początkowo sprawdzane są wszystkie możliwe ustawienia kąta RMA, a następnie naprzemiennie wykorzystywane są wszystkie najlepsze ustawienia zgłoszone przez odbiorniki. W wyniku symulacji odbiornik A jest w stanie z powodzeniem wykryć sygnał przez 32,08% czasu, odbiornik B jest w stanie z powodzeniem odebrać pożądany sygnał przez 33,96% czasu, a odbiornik C jest wolny od znaczących zakłóceń przez 79,25% czasu. Dodatkowo średnia moc zakłóceń obserwowana w odbiorniku C spada o około 8,6 dB. Dodatkowo sprawdzone zostało jeszcze jedno rozwiązanie, gdzie po sprawdzeniu wszystkich ustawień kąta RMA wybrane zostają na resztę symulacja dwa (zamiast trzech) ustawienia, dla których nadal spełnione są założone wymagania. W efekcie wybierane są na przemian dwa ustawienia: najlepsze dla detektora A oraz najlepsze dla odbiornika B. Decyzja ta została uzasadniona minimalizacją liczby rozwiązań, a tym samym liczby zmian w systemie, oraz tym, że najlepsze ustawienie dla odbiornika B jest tylko nieco gorszym rozwiązaniem dla odbiornika C. Uzyskano w ten sposób następujące rezultaty. Odbiornik A jest w stanie z powodzeniem wykryć sygnał przez 43,4% czasu, odbiornik B jest w stanie z powodzeniem odebrać pożądany sygnał przez 45,28% czasu, a odbiornik C jest wolny od znaczących zakłóceń przez 79,25% czasu. Średnia moc zakłóceń obserwowana w odbiorniku C spada o około 8,58 dB. Zatem udało się tym prostym zabiegiem znacznie poprawić jakoś obsługi odbiorników A i B, kosztem podniesienia średniej interferencji w odbiorniku C o 0,02 dB (średnia obserwowana interferencja nadal poniżej ustalonego dopuszczalnego progu).

4. PODSUMOWANIE

Przedstawione w artykule wyniki symulacji pokazują, że wykorzystanie RMA, nawet przy najprostszym mechanizmie sterowania, może przynieść korzyści w każdym z analizowanych przypadków użycia. Jednakże znacznie lepsze efekty można uzyskać implementując algorytm sterowania RMA i mechanizm dostarczania zainteresowanym odbiorcom informacji o aktualnym ustawieniu RMA. Daje to bardzo elastyczne rozwiązanie, którego działanie wymaga dalszej analizy. W szczególności należy wziąć pod uwagę charakter otoczenia, charakter źródła lub źródeł sygnału, potencjalne odbiorniki sygnału, ich położenie oraz cel użytkowania. Przedstawione koncepcje przede wszystkim stwarzają możliwość kierowania sygnału w miejsca wcześniej niedostępne – jednak należy pamiętać, że sytuacja ta może być zarówno pozytywna, jak i negatywna. Istnieje bowiem ryzyko wprowadzenia zakłóceń w miejscu, w którym dotychczas nie było nadmiernych zakłóceń. Problem ten pogłębia się zwłaszcza w odniesieniu do urządzeń nieświadomych istnienia i obecności RMA. Niemniej jednak pomysły przedstawione w tym artykule wydają się obiecujące pod względem potencjalnych korzyści dla pewnych zastosowań.

PODZIĘKOWANIE

Praca powstała w ramach projektu finansowanego przez Narodowe Centrum Nauki (nr 2021/43/B/ST7/01365).

LITERATURA

[1] Höyhtyä Markom. 2016. ”Spectrum Occupancy Measurements: A Survey and Use of Interference Maps”. IEEE Communications Surveys & Tutorials, 18 (4) : 2386-2414.

[2] Liaskos Christos. 2018. ”A New Wireless Communication Paradigm through Software-Controlled Metasurfaces”. IEEE Communications Magazine, 56 (9) : 162-169.

[3] Liu Yuanwei. 2021. ”Reconfigurable Intelligent Surfaces: Principles and Opportunities”. IEEE Communications Surveys & Tutorials, 23 (4) : 1546-1577.

[4] Renzo Marco. 2022. ”Communication Models for Reconfigurable Intelligent Surfaces: From Surface Electromagnetics to Wireless Networks Optimization”. Proceedings of the IEEE, 110 (9) : 1164-1209.

MODEL DWUWYMIAROWY DO SYMULACJI GĘSTOŚCI MOCY W OBSZARZE GŁOWY I OCZU DLA CZĘSTOTLIWOŚCI 26 GHz

TWO-DIMENSIONAL MODEL FOR POWER DENSITY SIMULATIONS
IN THE REGION OF HEAD AND EYES AT 26 GHz

Jarosław Kawecki[1]; Łukasz Januszkiewicz[2]; Paolo Di Barba[3]

[1] Politechnika Łódzka, Instytut Elektroniki, Łódź, jaroslaw.kawecki@p.lodz.pl
[2] Politechnika Łódzka, Instytut Elektroniki, Łódź, lukasz.januszkiewicz@p.lodz.pl
[3] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pawia, paolo.dibarba@unipv.it

DOI: 10.15199/59.2023.4.63

Streszczenie: **W artykule przedstawiono nowy model numeryczny umożliwiający obliczenie wartości średniej i maksymalnej gęstości mocy fali elektromagnetycznej w obszarze głowy i oczu dla częstotliwości 26 GHz. Jest to uproszczony model dwuwymiarowy typu 'ray-tracing', charakteryzujący się bardzo dużą wydajnością obliczeniową. Opracowanie modelu uproszczonego poprzedzone zostało zebraniem danych referencyjnych, które uzyskano za pomocą antropomorficznego modelu heterogenicznego i metody FDTD. Następnie przedstawiono opracowany model ze ścianami tłumiącymi i odbijającymi fale.**

Abstract: **The article presents a new numerical model for calculating the average and maximum value of electromagnetic wave power density in the area of the head and eyes for the frequency of 26 GHz. It is a simplified ray-tracing, two-dimensional model characterized by very high computational efficiency. The development of the simplified model was preceded by the collection of reference data obtained using the anthropomorphic heterogeneous model and the FDTD method. Then, the model with attenuating and reflecting walls was presented.**

Słowa kluczowe: **Modele numeryczne ciała człowieka, optymalizacja komputerowa, śledzenie promieni**

Keywords: **Numerical models of the human body, computer optimization, ray-tracing**

1. WSTĘP

We współczesnych systemach radiokomunikacji ruchomej wykorzystywane są coraz większe częstotliwości sięgające dziesiątek GHz. W systemach piątej generacji (5G) pasmo wysokie obejmuje częstotliwości powyżej 24 GHz. W przypadku tak dużych częstotliwości przeprowadzenie symulacji oddziaływania fali elektromagnetycznej na ciało człowieka wymaga wykorzystania znacznych zasobów obliczeniowych. Stosując metodę różnic skończonych w dziedzinie czasu (FDTD), która jest rekomendowaną metodą do analizy oddziaływania fal EM na ciało człowieka [1], wymagany jest podział symulowanego obiektu na komórki elementarne (woksele). Ich krawędzie nie mogą być dłuższe niż 1/10 długości najkrótszej fali w obiekcie [2]. W przypadku symulacji gęstości mocy we wnętrzu ciała dla częstotliwości 26 GHz stosowanej w systemach 5G, długość fali w ciele jest znacznie mniejsza od jego rozmiarów i wynosi około 1,6 mm, przyjmując średnią przenikalność elektryczną względną ciała równą 50. Konieczne jest zatem zastosowanie w tym przypadku wokseli o krawędziach nie większych niż 0,16 mm. W przypadku symulacji obszaru ciała skutkuje to bardzo dużymi wymaganiami dotyczącymi pamięci komputerowej (rzędu setek GB) oraz bardzo długim czasem symulacji.

W przypadku symulacji oddziaływania fal elektromagnetycznych z obiektami, które są wielokrotnie większe od długości fali stosowane są tzw. metody asymptotyczne, które przybliżają efekt rozchodzenia się fali za pomocą śledzenia promieni (ang. ray-tracing). Oddziaływanie fali z obiektem, które można zamodelować w tych metodach obejmuje zjawiska tłumienia fali w wolnej przestrzeni, tłumienia fali w ośrodku stratnym, odbicia od obiektów oraz ugięcia fali na przeszkodzie [3]. Metodą śledzenia promieni modelowano również zjawisko przysłonięcia łącza bezprzewodowego przez ciało człowieka. W publikacji [4] opisano model ciała człowieka umożliwiający modelowanie tego zjawiska w paśmie niskim 5G na częstotliwości 3,5 GHz za pomocą śledzenia promieni z uwzględnieniem dyfrakcji fali na krawędzi obiektu.

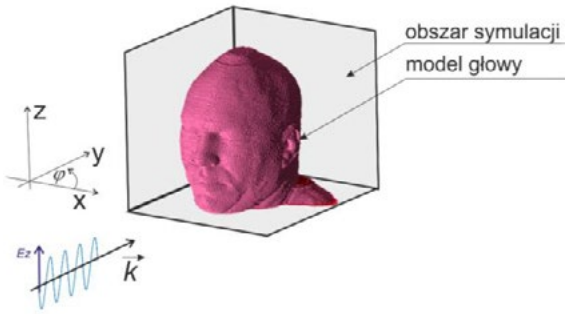
W artykule przedstawiono nowy model numeryczny umożliwiający obliczenie gęstości mocy fali elektromagnetycznej we wnętrzu ciała. Został on opracowany do wyznaczania średniej i maksymalnej gęstości mocy w obszarze głowy i oczu dla częstotliwości 26 GHz. Jest to uproszczony model dwuwymiarowy wykorzystujący śledzenie promieni ('ray-tracing'), charakteryzujący się bardzo dużą wydajnością obliczeniową. Opracowanie modelu uproszczonego poprzedzone zostało zebraniem danych referencyjnych, które uzyskano za pomocą antropomorficznego modelu heterogenicznego i metody FDTD.

2. SYMULACJE GĘSTOŚCI MOCY W OBSZARZE GŁOWY ZA POMOCĄ METODY FDTD

Zastosowanie metody FDTD do modelowania oddziaływania fali elektromagnetycznej na ciało w pasmie 26 GHz wymaga wykorzystania znacznych zasobów obliczeniowych. W celu zebrania danych referencyjnych przeprowadzono symulacje gęstości mocy w obszarze głowy poddanej ekspozycji na falę EM o częstotliwości 26 GHz, o polaryzacji liniowej pionowej i wartości skutecznej składowej elektrycznej wynoszącej E_z=1 V/m.

Geometria eksperymentu została przedstawiona na rys. 1. Fala EM rozchodzi się w płaszczyźnie *x-y*, a kierunek rozchodzenia się fali (kierunek wektora *k*) określa kąt φ=90°. Najistotniejszymi obszarami, na których skoncentrowano się w przeprowadzonym eksperymencie były oczy i mózg. Przenikalność elektryczna względna tkanek oczu wynosi dla tej częstotliwości od 14 (soczewka) do 31 (ciało szkliste) [5]. Przenikalność istoty szarej mózgu wynosi 22, a istoty białej 17. Z tego względu przyjęto że średnia wartość przenikalności elektrycznej tkanek głowy wynosi 20, a odpowiadająca jej długość fali to 2,6 mm. Przyjęto, że na potrzeby tej symulacji zastosowane mogą być woksele nie większe niż 0,25 mm.

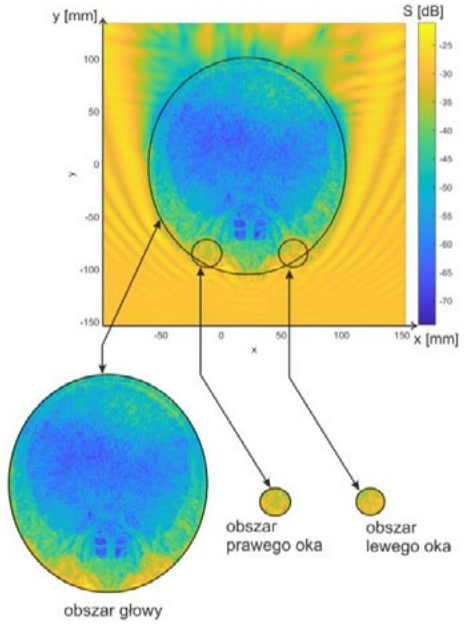
Do symulacji metodą FDTD wykorzystano program XFdtd firmy Remcom [6] oraz heterogeniczny model ciała człowieka dostępny w tym programie. Aby można było przeprowadzić symulację na komputerze dysponującym 48 GB pamięci operacyjnej, konieczne było ograniczenie oryginalnego modelu głowy do obszaru o wymiarach: 265 mm w kierunku osi *x*, 301 mm w kierunku osi *y* i 262 mm w kierunku osi *z*. Obszar ten został pokazany na rys. 1. Liczba wokseli tworzących model wyniosła 1 337 500 000. Rozmiar pamięci potrzebnej do uruchomienia symulacji to 39,6 GB. Symulacje przeprowadzono z wykorzystaniem procesora CPU Intel Xeon Silver 4110 taktowanego z częstotliwością 2,10 GHz z 8 rdzeniami fizycznymi i 16 rdzeniami logicznymi. Czas obliczeń wyniósł 9 godzin i 32 minuty. W tym przypadku nie można było wykorzystać dostępnego procesora graficznego GPU ponieważ dysponuje on jedynie 32 GB pamięci, którą można wykorzystać do symulacji.



Rys. 1. Obszar do którego ograniczono model głowy na potrzeby symulacji metodą FDTD

W wyniku symulacji uzyskano rozkład gęstości mocy, który na przekroju w płaszczyźnie poziomej na wysokości oczu przedstawiony został na rysunku 2.

Metoda FDTD pozwala uzyskać bardzo dokładne rozkłady gęstości mocy pola elektromagnetycznego wewnątrz ciała człowieka. W wielu zastosowaniach tak dokładne informacje nie są jednak potrzebne. W celu oszacowania, czy wrażliwe obszary ciała mogą być narażone na promieniowanie EM podczas projektowania ochron osobistych lub na potrzeby analizy złożonych warunków ekspozycyjnych wystarczające jest określenie średniej i maksymalnej gęstości mocy. W obszarze głowy takimi wrażliwymi organami są oczy. W celu łatwego wyznaczenia wartości średniej i maksymalnej gęstości mocy występującej w obszarze głowy został on przybliżony w postaci elipsy opisanej na przekroju poprzecznym głowy na wysokości oczu. Podobnie, aby wyznaczyć wartości maksymalne i średnie w obszarze oka te obszary również zostały przybliżone w postaci elips zawierających tkanki gałek ocznych. Na rys. 2 pokazano elipsę przybliżającą obszar głowy oraz elipsy przybliżające obszar gałki ocznej. Promień elipsy przybliżający głowę równoległy do osi *x* wynosi r_{gx}=82 mm a równoległy do osi *y*, r_{gy}= 103 mm. Promień elipsy dla oka prawego równoległy do osi *x* r_{opx}= 12,5 mm, a równoległy do osi *y* r_{opy}=13 mm. Promień elipsy dla oka lewego równoległy do osi *x* r_{olx}= 12 mm, a równoległy do osi *y* r_{oly}=13 mm. Dla tych obszarów, na podstawie wyników symulacji FDTD obliczono średnią i maksymalną wartość gęstości mocy. Wartość maksymalna dla obszaru głowy wyniosła S_{mg}=-21,1 dBW/m^2, wartość średnia dla obszaru głowy wyniosła S_{sg}= -51,1 dBW/m^2. Wartości dla prawego oka wyniosły: maksymalna S_{mop}=-23,5 dBW/m^2, średnia S_{sop}= -32,5 dBW/m^2. Dla oka lewego wartość maksymalna wyniosła S_{mol}= -23,6 dBW/m^2, wartość średnia S_{sol}= -33,4 dBW/m^2. Opisany dalej model uproszczony będzie miał za zadanie odwzorowanie tych samych wartości w rozważanych obszarach dla identycznych parametrów fali elektromagnetycznej.

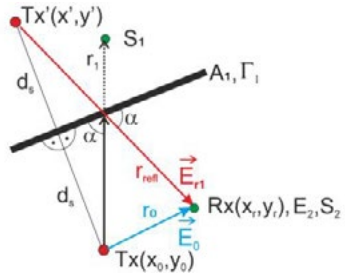


Rys. 2. Rozkład gęstości mocy S na przekroju poprzecznym głowy, na wysokości oczu. Skala logarytmiczna odniesiona do 1W/m^2, φ=90°

3. IDENTYFIKACJA PARAMETRÓW MODELU UPROSZCZONEGO

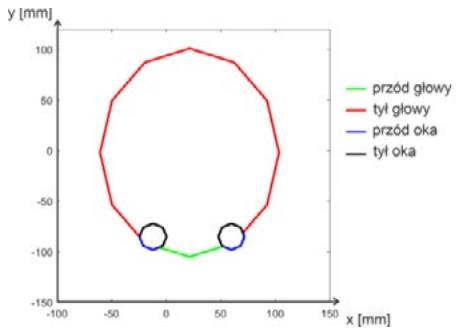
Jedną z najprostszych metod wykorzystywanych do symulacji elektromagnetycznych obiektów wielokrotnie większych od długości fali jest stosowany w symulacjach bazujących na śledzeniu promieni ('ray-tracing') model wielościenny (ang. multi-wall) stosowany we wnętrzach budynków. W najprostszym ujęciu model wielościenny jest modelem dwuwymiarowym w którym rozpatruje się rozchodzenie się fali w płaszczyźnie poziomej [7]. Pomijane są wtedy oddziaływania z przegrodami równoległymi do kierunku rozchodzenia się fali. W przedstawionym modelu ciała człowieka zastosowano analogiczne

podejście uwzględniając dodatkowo odbicie fali od ściany. Schematycznie jego zasadę działania przedstawiono na rys. 3. Gęstość mocy S_1 w punkcie oddalonym o r_1 od nadajnika jest obliczana analogicznie jak w modelu wielościennym. Jej wartość wynika z tłumienia wolnej przestrzeni dla odległości r_1 oraz tłumienia ściany o wartości A_1. W przypadku gęstości mocy S_2 w punkcie R_x, obliczana jest ona z natężenia pola elektrycznego E_2 które jest sumą dwóch składowych: składowej rozchodzącej się w wolnej przestrzeni E_0 i składowej odbitej od pierwszej ściany E_{r1}. Dla każdego punktu, w którym obliczana jest gęstość mocy w modelu i dla każdej ściany w omawianym modelu sprawdzane jest z warunku Snelliusa czy dana ściana powoduje odbicie fali, w rezultacie którego fala odbita może dotrzeć do tego punktu. Uwzględniane jest jednokrotne odbicie. W tym modelu dla uproszczenia przyjęto że zarówno współczynnik tłumienia ściany A jak i jej współczynnik odbicia Γ nie zależą od kąta padania fali.



Rys. 3. Promienie w modelu wielościennym z jednokrotnym odbiciem

Opracowany model do symulacji oddziaływania fali elektromagnetycznej na ciało odwzorowuje przekrój poprzeczny głowy na wysokości środka gałek ocznych. Głowa oraz oczy odwzorowane są za pomocą ścian ułożonych na elipsach ograniczających obszary głowy i oczu, które pokazano na rys. 2. Elipsy ograniczające te obszary przybliżone są w modelu za pomocą wielokątów, których boki charakteryzuje określone tłumienie A oraz współczynnik odbicia Γ. Na rys. 4 pokazano model przybliżający obszar głowy za pomocą 12 ścian, a gałkę oczną za pomocą 8 ścian. Różnymi kolorami zaznaczone zostały ściany o tym samym tłumieniu i współczynniku odbicia. Założono 4 różne tłumienia ścian tworzących model dla różnych obszarów: A_1 przód głowy, A_2 tył głowy, A_3 przód oka i A_4 tył oka oraz analogicznie dla współczynników odbicia: Γ_1 do Γ_4.



Rys. 4. Model 2 wymiarowy głowy i oczu złożony z 12 ścian odwzorowujących głowę i 8 ścian odwzorowujących gałki oczne

Wartości tłumienia ścian tworzących model oraz współczynniki odbicia ścian zostały dobrane z wykorzystaniem algorytmu automatycznej optymalizacji EStra w wariancie najniższego rzędu [8], który był wcześniej stosowany przez autorów do optymalizacji zagadnień elektromagnetycznych [9,10]. W tym przypadku funkcją celu f, która podlegała minimalizacji była maksymalna różnica pomiędzy wartościami średnimi i maksymalnymi gęstości mocy uzyskanymi z modelu odniesienia (FDTD) i modelu wielościennego:

$$f(A_1 \dots A_4, \Gamma_1 \dots \Gamma_4) = \max_{n=1,3}\{|S_{ref\ max}(n) - S_{simp\ max}(n, A_1 \dots A_4, \Gamma_1 \dots \Gamma_4)|, |S_{ref\ avg}(n) - S_{simp\ avg}(n, A_1 \dots A_4, \Gamma_1 \dots \Gamma_4)|\} \quad (1)$$

gdzie:
$S_{ref\ max}$ — maksymalne wartości gęstości mocy uzyskane z modelu referencyjnego (W/m^2)
$S_{simp\ max}$ — maksymalne wartości gęstości mocy uzyskane z modelu uproszczonego (W/m^2)
$S_{ref\ avg}$— średnie wartości gęstości mocy uzyskane z modelu referencyjnego (W/m^2)
$S_{simp\ avg}$ — średnie wartości gęstości mocy uzyskane z modelu uproszczonego (W/m^2)
n — numer rozważanego obszaru: 1 – głowa, 2 – prawe oko, 3 – lewe oko.

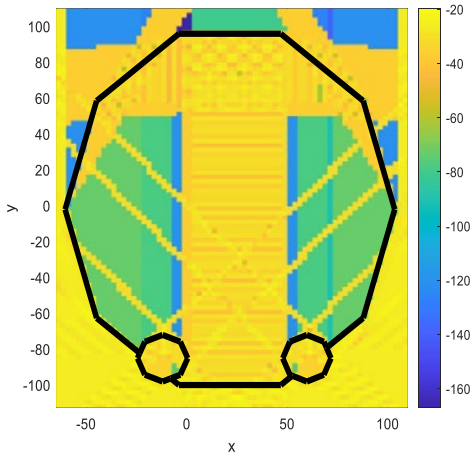
W tym ujęciu zastosowana metoda identyfikacji parametrów modelu jest optymalizacją wielokryterialną metodą Min-Max. W przypadku algorytmu EStra konieczne jest przyjęcie wartości początkowych poszukiwanych parametrów. Założono, że początkowe wartości tłumienia ścian będą wynosiły 10 dB, a ich współczynniki odbicia 0. Ze względu na to że zastosowano algorytm stochastyczny to przyjęcie konkretnych wartości początkowych poszukiwanych paramentów nie wpływa znacząco na jakość otrzymanego rozwiązania.

Wartość funkcji celu wyznaczana jest na podstawie wartości średnich i maksymalnych gęstości mocy odpowiednich obszarów. Symulacje przeprowadzono dla 6390 punktów usytuowanych wewnątrz prostokątnego obszaru (obejmującego głowę i oczy) o wymiarach 175 mm na 222 mm. Współrzędne x punktów zmieniały się co 2,5 mm w przedziale od -65 mm do 110 mm. Współrzędne y punktów zmieniały się co 2,5 mm w przedziale od -112, do 110 mm.

4. WYNIKI UZYSKANE ZA POMOCĄ MODELU WIELOŚCIENNEGO

Procedurę poszukiwania parametrów modelu przeprowadzono dla wielu wariantów różniących się liczbą ścian odwzorowujących obszar oczu i głowy. Najlepsze wyniki uzyskano dla wariantu z 10 ścianami odwzorowującymi obszar głowy i 8 ścianami odwzorowującymi obszar oczu. W tym przypadku najmniejszą wartość funkcji celu f=0,59 dB uzyskano po 54 iteracjach. Wartości zidentyfikowanych dla tego przypadku parametrów modelu wyniosły odpowiednio: A_1=7,02, A_2=47,51, A_3=9,05, A_4=43,36, Γ_1=0,01, Γ_2=-0,29, Γ_3=0,52, Γ_4=-0,51. Na rys.5 przedstawiono wyniki symulacji gęstości mocy wyzyskane dla tego modelu.

W tabeli 1 zostały przedstawione maksymalne i średnie wartości gęstości mocy wyznaczone dla obszarów głowy i oka za pomocą modelu ze ścianami tłumiącymi i odbijającymi. Dotyczą one przypadku, w którym osiągnięto najmniejszą wartość funkcji celu. Te wyniki porównane zostały z odpowiednimi parametrami uzyskanymi dla modelu referencyjnego. Największa różnica (0,59 dB) występuje w przypadku wartości średniej dla obszaru głowy. W przypadku pozostałych wartości średnich i maksymalnych ich różnice nie przekraczają tej wartości.



Rys. 5. Wyniki symulacji gęstości mocy wyzyskane dla modelu wielościennego głowy

Tab. 1. Maksymalne i średnie gęstości mocy wyznaczone dla obszarów głowy i oka za pomocą zoptymalizowanego modelu ze ścianami tłumiącymi i odbijającymi oraz modelu referencyjnego.

Gęstość mocy	*Model FDTD [dBW/m²]*	*Model uproszczony [dBW/m²]*
Maksimum, głowa	-21,16	-21,64
Maksimum, prawe oko	-23,65	-23,07
Maksimum, lewe oko	-23,45	-22,94
Średnia, głowa	-51,13	-51,07
Średnia, prawe oko	-33,44	-34,03
Średnia, lewe oko	-32,53	-32,95

Opracowany model umożliwia bardzo szybkie uzyskanie wyników. Dla 6390 punktów stanowiących obszar obliczeń czas symulacji wynosi około 48 sekund na komputerze wyposażonym w procesor CPU Intel Core i7-9850H taktowany z częstotliwością 2.60GHz. W czasie symulacji środowisko Matlab, w którym zaimplementowany był model wykorzystywało 1,9 GB pamięci operacyjnej RAM.

5. WNIOSKI

Przedstawiony model wielościenny został opracowany w celu wyznaczania wartości średniej i maksymalnej gęstości mocy w obszarze głowy. Uzyskane z jego pomocą wyniki nie odbiegają od wyników symulacji uzyskanych za pomocą metody FDTD o więcej niż 0,6 dB. Jest to wystarczająco dobre przybliżenie w przypadku oszacowywania potencjalnych zagrożeń dla organów wrażliwych. W odróżnieniu jednak od modelu referencyjnego w metodzie FDTD, czas potrzebny do uzyskania wyników jest ponad 700 krotnie krótszy. Również wymaganie dotyczące pamięci operacyjnej jest zdecydowanie mniejsze niż w przypadku modelu odniesienia. Ze względu na złożoną budowę wewnętrzną ciała opracowany model nie pozwala na uzyskanie rozkładów gęstości mocy w tkankach głowy, które byłyby zbliżone do wyników symulacji metodą FDTD. Jego zastosowanie obejmuje jedynie wyznaczanie wartości średniej i maksymalnej w założonych obszarach ciała.

Przedstawiony powyżej dwuwymiarowy model ze ścianami tłumiącymi i jednokrotnym odbiciem został opracowany przy założeniu jednej wartości kąta padania fali φ=90°. Umożliwia on dobre przybliżenie wartości maksymalnych i średnich gęstości mocy jedynie dla zbliżonych warunków ekspozycji ciała na falę. Dalsze badania będą miały na celu zweryfikowanie możliwości opracowania analogicznego modelu dla szerszego zakresu kątów.

LITERATURA

[1] "IEEE Recommended Practice for Measurements and Computations of Electric, Magnetic, and Electromagnetic Fields with Respect to Human Exposure to Such Fields, 0 Hz to 300 GHz". 2021. *IEEE Std C95.3-2021*

[2] Taflove Allen, Susan C. Hagness. 2005. "Computational Electrodynamics: The Finite-Difference Time-Domain Method". Artech House: Norwood, MA, USA.

[3] Xueda Niu, Haibo He, Ming Jin. 2021. "Application of Ray-Tracing Method in Electromagnetic Numerical Simulation Algorithm,". *2021 International Applied Computational Electromagnetics Society (ACES-China) Symposium,* Chengdu, China,

[4] Januszkiewicz Łukasz. 2018. „Model for Ray-Based UTD Simulations of the Human Body Shadowing Effect in 5G Wireless Systems". *International Journal of Antennas and Propagation*. vol. 2018.

[5] https://itis.swiss/virtual-population/tissue-properties/database/dielectric-properties/

[6] Reference Manual-XFdtd. 2023. *https://support.remcom.com/xfdtd/reference.html*

[7] Perez Fontan Fernando, Perfecto Marino Espineira. 2008. „Modeling the wireless propagation channel : a simulation approach with MATLAB", *John Wiley & Sons Ltd.*

[8] Di Barba Paolo. 2010. „Multiobjective shape design in electricity and magnetism". *Springer*

[9] Di Barba Paolo, Maria Evelina Mognaschi, Ryszard Palka, Antonio Savini. 2009. "Optimization of the MIT Field Exciter by a Multiobjective Design". *IEEE Transactions on Magnetics,* 45-3:1530-1533.

[10] Januszkiewicz Łukasz, Paolo Di-Barba, Sławomir Hausman. 2016. „Automatyczna identyfikacja parametrów uproszczonego modelu ciała człowieka za pomocą algorytmu ewolucyjnego", *Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne*, ,6: 552-555.

MAPY ŚRODOWISKA RADIOWEGO BAZUJĄCE NA PEM W SIECIACH FANET

PEM-BASED RADIO ENVIRONMENT MAPS IN FANETS

Michał Kryk[1]; Krzysztof Malon[2]; Jan M. Kelner[3]

[1,2,3] Wojskowa Akademia Techniczna im. Jarosława Dąbrowskiego, Wydział Elektroniki, Instytut Systemów Łączności
ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warszawa, {[1]michał.kryk, [2]krzysztof.malon, [3]jan.kelner}@wat.edu.pl

DOI: 10.15199/59.2023.4.64

Streszczenie: **Mapa środowiska radiowego to pomocne narzędzie, które jest obecnie wykorzystywane do planowania sieci radiowych i zarządzania nimi. Znajdują one zastosowanie np. w sieciach radia kognitywnego lub mobilnych sieciach doraźnych. W przypadku sieci doraźnych z węzłami rozmieszczonymi na bezzałogowych statkach powietrznych, wykorzystanie dwuwymiarowych map jest niewystarczające. Ten artykuł koncentruje się na generowaniu trójwymiarowej mapy, która jest zbiorem map dwuwymiarowych zdefiniowanych dla dyskretnych wysokości nad poziomem gruntu lub morza. Każda taka mapa jest wyznaczana za pomocą metody równań parabolicznych, która pozwala określić tłumienia dróg propagacyjnych dla wielu profili terenu rozchodzących się promieniście od analizowanego węzła sieci. W artykule przedstawiono opracowany algorytm oraz przykładowe wyniki dla rozległego obszaru.**

Abstract: **A radio environment map is a helpful tool currently used to plan and manage radio networks. They are used, e.g., in cognitive radio networks or mobile ad-hoc networks. In the case of flying ad-hoc networks, the use of two-dimensional maps is insufficient. This paper focuses on generating a three-dimensional map as a set of two-dimensional maps defined for discrete heights above ground or sea level. Each map is determined using the parabolic equation method, which allows for determining the path losses for many terrain profiles outgoing radially from the analyzed transmitting node. The paper presents the developed algorithm and exemplary results for a large area.**

Słowa kluczowe: **sieć FANET, mapa środowiska radiowego (REM), metoda równania parabolicznego (PEM), propagacja fal, tłumienie propagacyjne.**

Keywords: **flying ad-hoc network (FANET), parabolic equation method (PEM), path loss, radio environment map (REM), wave propagation.**

1. WSTĘP

Mapa środowiska radiowego REM (*Radio Environment Map*) jest istotnym elementem architektury nowoczesnego systemu łączności bezprzewodowej, wykorzystywanym głównie do zarządzania zasobami radiowymi [2,11]. Wdrażanie algorytmów sieci radia kognitywnego lub dynamicznego zarządzania widmem spowodowało wzrost znaczenia REM [8,9]. Odpowiednia alokacja zasobów widmowych dla kompatybilnej pracy sieci w środowisku elektromagnetycznym oraz oszacowania zasięgu węzłów ma istotne znaczenie dla przestrzennego rozmieszczenia węzłów sieci ruchomej w terenie o określonym ukształtowaniu czy stopniu urbanizacji [12]. Stąd też, REM są integralną częścią współczesnych mobilnych sieci doraźnych MANET (*Mobile Ad hoc NETwork*), w tym taktycznych MANET wykorzystywanych w wojsku [3,14].

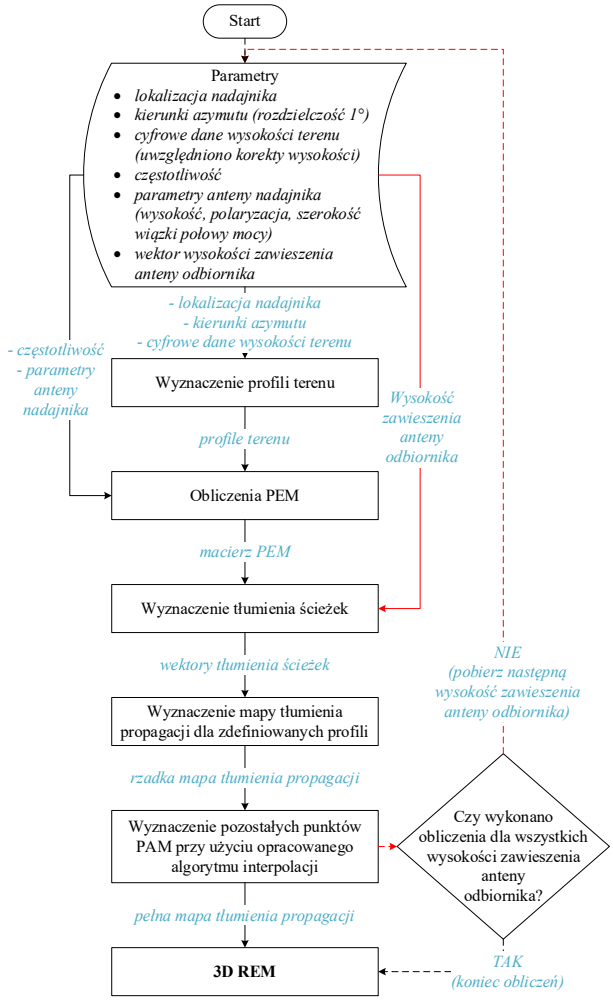
Analizując specyfikację i standard 5G New Radio, można dostrzec wiele nowoczesnych technologii, które mogą współpracować lub być oparte na REM. Wśród nich wyróżnić można wspomniane zarządzanie zasobami radiowymi, dynamiczne zarządzanie widmem czy sieci radia kognitywnego związane z algorytmami sztucznej inteligencji stosowanymi w 5G [4,13]. Idea REM może znaleźć zastosowanie także w stacji bazowej 5G (*gNodeB*) bazującej na technologii massive MIMO (*Multiple-Input-Multiple-Output*), która wykorzystuje algorytmy dynamicznie zmiennego formowania wiązki, tzw. *grid of beams* (siatka wiązek) lub *eigen-based* [1,10].

Do generowania REM stosuje się metody bezpośrednie lub pośrednie, tj. oparte odpowiednio na pomiarach lub modelach propagacyjnych [2,11]. W przypadku sieci działających na niewielkim obszarze metody bezpośrednie mogą się dobrze sprawdzić. Jednak w przypadku sieci działających na dużym obszarze, bardziej uniwersalne są metody pośrednie. Dla rozległego obszaru proponuje się zastosowanie metody REM opartej na metodzie równania parabolicznego PEM (*Parabolic Equation Method*). Rozwiązanie to pozwala wiernie odzwierciedlić wpływ ukształtowania terenu [5–7].

Dwuwymiarowe (2D) REM są wystarczające do analizy sieci radiowych dla lądowych sieci MANET, radia kognitywnego lub taktycznych. W przypadku sieci komórkowych LTE (*Long Term Evolution*) lub 5G, 2D REM są odpowiednie dla terenów płaskich wiejskich lub podmiejskich. W obszarach miejskich i metropolitalnych z wielopiętrowymi budynkami pomocne może być użycie trójwymiarowych (3D) REM. W przypadku sieci doraźnych z węzłami rozmieszczonymi na bezzałogowych statkach powietrznych UAV (*Unmanned Aerial Vehicle*), czyli tzw. sieci FANET (*Flying Ad-hoc NETwork*) niezbędne jest zastosowanie 3D REM [15]. Takich rozwiązań jest obecnie niewiele. W tym artykule zaproponowano 3D REM oparty na PEM, który może być wykorzystany do oceny zasięgu UAV względem naziemnej stacji kontroli, przekaźnika umieszczonego na UAV z węzłami naziemnymi lub zasięgu UAV w roju (klasyczny FANET). Proponowane rozwiązanie 3D bazuje na 2D REM przedstawionym w [5–7]. W tym przypadku 3D REM uzyskuje się przez wykonanie obliczeń dla dyskretnego zestawu wysokości nad poziomem gruntu.

2. ALGORYTM GENEROWANIA REM

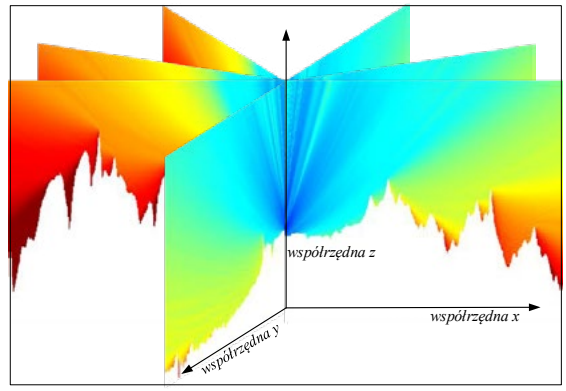
3D REM rozumiany jest jako zestaw map tłumienia propagacyjnego wyznaczonego dla dyskretnych wartości wysokości zawieszenia anteny odbiorczej. Proces tworzenia mapy przedstawiono na Rys. 1.



Rys. 1. Algorytm 3D REM

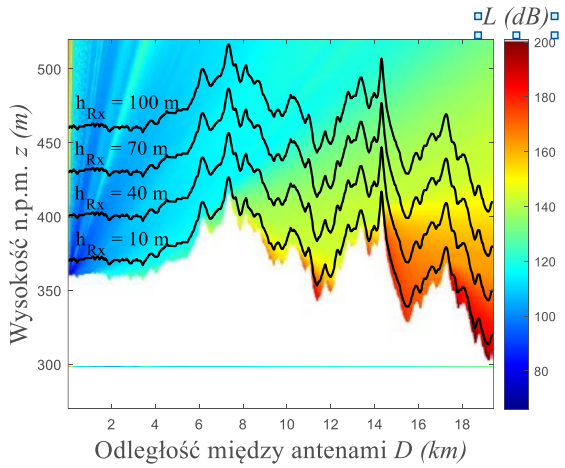
W górnej części schematu przedstawiono parametry i dane wejściowe niezbędne dla zaproponowanego algorytmu. Najpierw obliczane są profile terenu dla określonej lokalizacji nadajnika i azymutu na podstawie cyfrowych danych wysokości terenu. Następnie na podstawie wyznaczonych profili, dodatkowych parametrów anteny nadawczej oraz częstotliwości przeprowadzane są obliczenia PEM. Rys. 2 ilustruje widok 3D przykładowego PEM dla kilku azymutów względem lokalizacji nadajnika.

Na podstawie macierzy PEM i wysokości zawieszenia anteny odbiornika możliwe jest utworzenie tzw. rzadkiej macierzy z wartościami tłumienia drogi propagacyjnej na zdefiniowanych azymutach. Aby uzyskać pełną macierz tłumień wokół analizowanej pozycji nadajnika, należy wyznaczyć tłumienie również w obszarach poza punktami odpowiadającymi zdefiniowanym profilom terenu. W tym celu zaproponowano dedykowany algorytm interpolacji opisany w [7].



Rys. 2. Przykładowe 3D REM dla wybranych azymutów

Finalna REM dla określonej wysokości zawieszenia anteny odbiornika jest jedną z warstw 3D REM. W celu wyznaczenia kolejnych warstw tej mapy konieczne jest wykonanie obliczeń dla kolejnych wysokości zawieszenia anteny odbiorczej w pętli zaznaczonej na Rys. 1. Należy zauważyć, że warstwy te można otrzymać bez dodatkowego nakładu obliczeniowego dotyczącego wyznaczania profili i obliczeń PEM. Proces ten oparty jest na macierzy PEM, co zilustrowano na Rys. 3. Zastosowany na rysunku symbol L oznacza tłumienie propagacyjne wyrażone w dB.



Rys. 3. Przykładowe PEM dla wybranego azymutu i ukształtowania terenu z zaznaczonymi różnymi wysokościami zawieszenia anten odbiorczych

Warto również zwrócić uwagę na to, jak definiowana jest wysokość zawieszenia anteny odbiorczej. Ogólnie rzecz biorąc, mogą to być bezwzględne wysokości nad poziomem morza lub określone w dowolnym punkcie względem ziemi. W artykule zaproponowano analizę tłumienia na względnych wysokościach zawieszenia anteny z uwzględnieniem profilu wysokościowego terenu. Takie rozwiązanie można zastosować w scenariuszu wykorzystania map tłumienia np. do analizy zasięgu radiowego urządzeń komunikacyjnych zamontowanych na UAV.

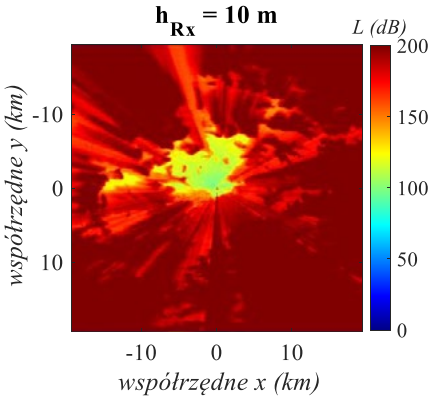
3. PRZYKŁADOWE 3D REM

Przykładowe REM zostały wygenerowane dla węzła sieci (nadajnika) położonego w punkcie o współrzędnych (50,4408ºN; 19,4911ºE). Symulacje te przeprowadzono dla czterech wysokości zawieszenia anten odbiorczych (pułapu lotu UAV), które przedstawiono na Rys. 3, tj. 10, 40, 70 oraz 100 m. Główne parametry zastosowane w badaniu zawiera .

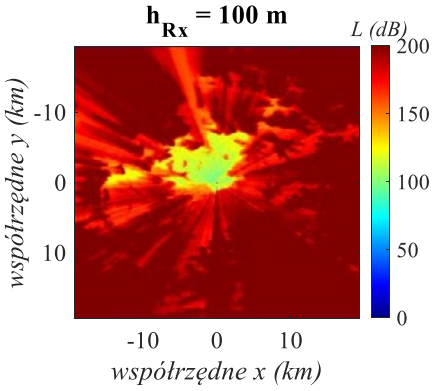
Zgodnie z algorytmem przedstawionym na Rys. 1 wygenerowane zostały cztery REM dla analizowanych wysokości zawieszenia anten odbiorczych przy założeniu maksymalnego zasięgu radiowego 20 km. Dwa REM dla skrajnych wysokości zawieszenia anteny pokazano na Rys. 4 i Rys. 5.

Tab. 1. Podstawowe parametry użyte w badaniach

Parametr	Wartość	Jednostka
Maksymalny zasięg	20	km
Rozdzielczość profili terenu	30	m
Maksymalny pułap lotu	150	m
Wysokość zawieszenia anteny nadawczej	1,5	m
Częstotliwość nośna sygnału	2,4	GHz
Szerokość wiązki głównej anteny nadawczej w płaszczyźnie pionowej	45	deg
Polaryzacja anten	pionowa	
Rodzaj refrakcji	normalna	

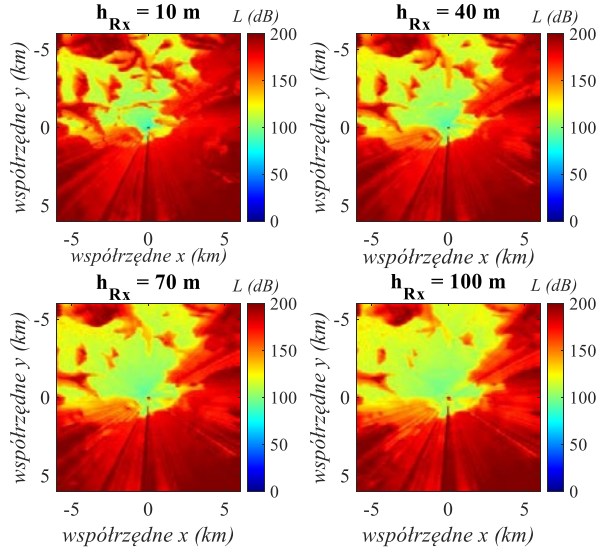


Rys. 4. REM dla wysokości zawieszenia anteny odbiorczej równej 10 m oraz zasięgu 20 km



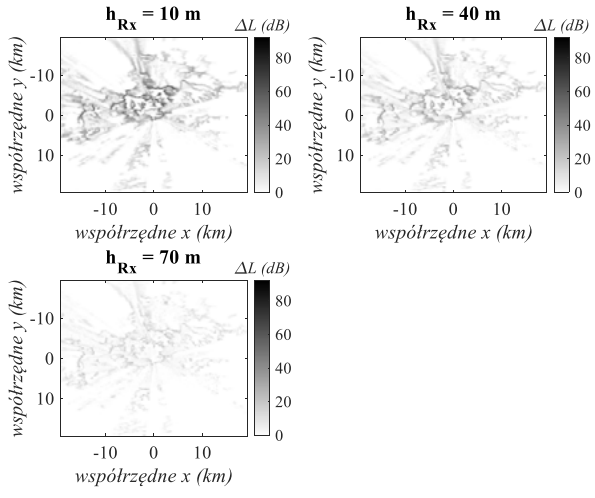
Rys. 5. REM dla wysokości zawieszenia anteny odbiorczej równej 100 m oraz zasięgu 20 km

W przypadku obu wygenerowanych map na znaczącej części obszaru występują duże wartości tłumień propagacyjnych (równych lub powyżej 200 dB). Wynika to z niskiego (około 2 m) montażu anteny nadawczej, co miało odzwierciedlać warunki pilota UAV. Dlatego też, na Rys. 6, promień analizy profilu zmniejszono do 6 km. W tym przypadku pokazane są REM dla wszystkich analizowanych wysokości zawieszenia anten odbiorczych. W tym przypadku różnice pomiędzy REM dla różnych wysokości zawieszenia anten są bardziej widoczne.



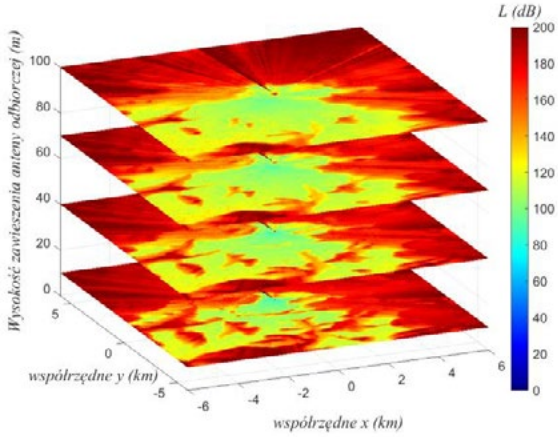
Rys. 6. REM dla wysokości zawieszenia anteny odbiorczej równej 10, 30, 70 i 100 m oraz zasięgu 6 km

Aby lepiej zilustrować te różnice, na Rys. 7 zilustrowano tzw. różnicowe REM. W tym przypadku różnice wartości tłumienia (ΔL) wykonano pomiędzy REM na danej wysokości i na maksymalnej analizowanej (odniesienia), czyli 100 m. Z tego też powodu zilustrowano tylko trzy mapy różnicowe. Połączenie wszystkich REM daje 3D REM, co przedstawiono na Rys. 8. Taki REM można również zilustrować w nieco innej formie za pomocą konturów, jak pokazano na Rys. 9.

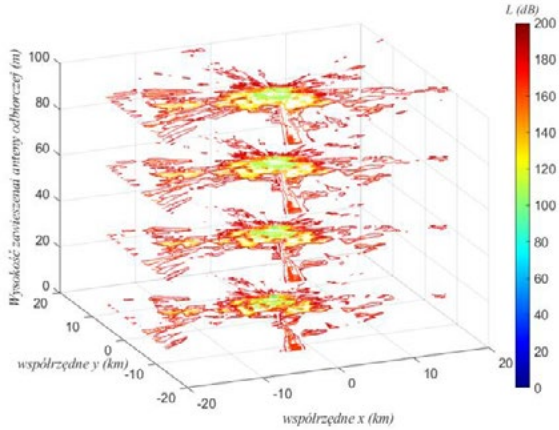


Rys. 7. Różnicowe REM dla wysokości zawieszenia anteny odbiorczej równej 10, 30 i 70 oraz zasięgu 6 km

W tym artykule zilustrowano przykładową 3D REM dla czterech wybranych wysokości zawieszenia anteny odbiorczej nad poziomem gruntu. Podczas generowania REM, w zależności od potrzeb, należy określić zakres wysokości oraz krok zmiany wysokości, dla którego obliczany jest 2D REM (w analizowanym przypadku wynosi on 30 m). Warto podkreślić, że REM można również wykonać na pewnych wysokościach względem poziomu morza.



Rys. 8. 3D REM dla dyskretnych wysokości zawieszenia anteny odbiorczej oraz zasięgu 6 km



Rys. 9. 3D REM dla dyskretnych wysokości zawieszenia anteny odbiorczej oraz zasięgu 6 km

3D REM mogą być wykorzystywane do planowania i zarządzania misjami pojedynczego UAV, roju dronów lub całej sieci FANET. Te autonomiczne platformy mogą latać na różnych wysokościach, co wymaga użycia map 3D.

4. PODSUMOWANIE

W artykule przedstawiono algorytm generowania 3D REM, który może być wykorzystany do wyznaczania tłumienia i zasięgu węzłów sieci mobilnych na terenach miejskich (z uwzględnieniem specyfiki wysokich budynków), planowania misji UAV czy zarządzania sieciami FANET. Proponowane rozwiązanie wykorzystuje dwuwymiarowe REM wyznaczone na podstawie obliczeń bazujących na metodzie równania parabolicznego i algorytmu aproksymacji opisanego w [7]. Takie podejście uwzględnia ukształtowanie terenu i może tworzyć REM na dużych obszarach. W artykule przedstawiono uproszczony opis algorytmu 3D REM oraz przykładowe wyniki uzyskane dla wybranej lokalizacji.

LITERATURA

[1] Chikha Wassim Ben, M. Masson, Z. Altman, S.B. Jemaa. 2023. "Radio environment map based inter-cell interference coordination for massive-MIMO systems". *IEEE Transactions on Mobile Computing*, 1–12.

[2] Dagres Ioannis *et al.* 2011. "Radio environmental maps: Information models and reference model". ICT-248351 FARAMIR, Document no. D4.1.

[3] Kaniewski Piotr, J. Romanik, E. Golan, K. Zubel. 2021. "Spectrum awareness for cognitive radios supported by radio environment maps: Zonal approach". *Applied Sciences*, 11 (7) : 7.

[4] Kliks Adrian, L. Goratti, T. Chen. 2016. "REM: Revisiting a cognitive tool for virtualized 5G networks". *2016 23rd International Conference on Telecommunications (ICT)*, 1–5.

[5] Kryk Michał. 2022. „Ocena efektywności generacji map tłumienia propagacyjnego". *Przegląd Telekomunikacyjny – Wiadomości Telekomunikacyjne*, 95 (4) : 429-432.

[6] Kryk Michał, K. Malon, J.M. Kelner. 2021. "Algorithm of creating propagation attenuation maps based on parabolic equation method". *2021 Signal Processing Symposium (SPSympo)*, 1–5.

[7] Kryk Michał, K. Malon, J.M. Kelner. 2022. "Propagation attenuation maps based on parabolic equation method". *Sensors*, 22 (11) : 4063.

[8] Malon Krzysztof, P. Skokowski, J. Łopatka, 2017, "The use of a wireless sensor network to monitor the spectrum in urban areas". Proceedings vol. 10418, XI Conference on Reconnaissance and Electronic Warfare Systems, 1041809.

[9] Malon Krzysztof, P. Skokowski, J. Łopatka, 2018, "Optimization of the MANET topology in urban area using redundant relay points". 2018 International Conference on Military Communications and Information Systems (ICMCIS), 1–4.

[10] Masson Marie Yolande Renée. 2023. "Self-organizing functionalities in 5G virtual radio access network", PhD thesis. Université d'Avignon.

[11] Pesko Marko, T. Javornik, A. Košir, M. Štular, M. Mohorčič, "Radio environment maps: The survey of construction methods". *KSII Transactions on Internet and Information Systems*, 8 (11) : 3789–3809.

[12] Skokowski Paweł, K. Malon, J. Łopatka, 2022, "Building the electromagnetic situation awareness in MANET cognitive radio networks for urban areas". *Sensors*, 22 (3) : 716.

[13] Suarez Rodriguez Antonio Cristo, N. Haider, Y. He, E. Dutkiewicz. 2020. "Network optimisation in 5G networks: A radio environment map approach". *IEEE Transactions on Vehicular Technology*, 69 (10) : 12043–12057.

[14] Suchański Marek, P. Kaniewski, J. Romanik, E. Golan, K. Zubel. 2022. "Radio environment maps for military cognitive networks: Density of small-scale sensor network vs. map quality". *EURASIP Journal on Wireless Communications and Networking*, 2020 (1) : 189.

[15] Theile Mirco, H. Bayerlein, R. Nai, D. Gesbert, M. Caccamo. 2021. "UAV path planning using global and local map information with deep reinforcement learning". *2021 20th International Conference on Advanced Robotics (ICAR)*, 539–546.

UPROSZCZONE MODELE CHARAKTERYSTYK ANTENOWYCH W ZASTOSOWANIU DO OCENY INTERFERENCJI MIĘDZYWIĄZKOWYCH

SIMPLIFIED MODELS OF ANTENNA PATTERNS TO INTER-BEAM INTERFERENCE ASSESSMENT

Jarosław Wojtuń[1]; Cezary Ziółkowski[2]; Jan M. Kelner[3]
[1,2,3] Wojskowa Akademia Techniczna im. Jarosława Dąbrowskiego, Wydział Elektroniki, Instytut Systemów Łączności
ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warszawa, {[1]jaroslaw.wojtun, [2]cezary.ziolkowski, [3]jan.kelner}@wat.edu.pl

DOI: 10.15199/59.2023.4.65

Streszczenie: **Występowanie międzywiązkowych zakłóceń w odbieranym sygnale jest problemem, który ogranicza możliwości technologii massive MIMO (multiple-input, multiple-output). Zmianom kierunku wiązek w antenach rzeczywistych towarzyszą zmiany kształtu ich charakterystyk i parametrów co utrudnia ocenę tych zakłóceń. W tym artykule, prezentowana jest metodyka oceny zakłóceń SIR (signal-to-interference ratio) pochodzących od sąsiednich wiązek, która wykorzystuje modyfikacje prostych modeli charakterystyk wiązek antenowych. Ocenę dokładności metodyki przeprowadzono na podstawie wyników badań symulacyjnych z wykorzystaniem wieloelipsoidalnego modelu propagacji.**

Abstract: **The occurrence of inter-beam interference in the received signal is a problem that limits the possibilities of the massive MIMO technology. In real antennas, the changes in the shape of their characteristics and parameters occur together with the direction changes of beams. This makes it difficult to assess this interference. This paper presents the methodology for assessing signal-to-interference ratio (SIR) from adjacent beams. This methodology is based on simple models of antenna beam patterns. The accuracy of the methodology is assessed based on the simulation study results using a multi-ellipsoidal propagation model.**

Słowa kluczowe**: modele charakterystyk antenowych, technologia massive MIMO, wieloelipsoidalny model propagacji, zakłócenia międzywiązkowe.**

Keywords**: inter-beam interference, massive MIMO technology, models of antenna patterns, multi-ellipsoidal propagation model.**

1. WSTĘP

Wykorzystanie w bezprzewodowych systemach 5G zakresu fal milimetrowych daje możliwość wprowadzenie nowych technologii, które w istotnym stopniu przyczyniają się do zwiększenia efektywność zarządzania dostępnymi zasobami radiowymi [6]. Massive MIMO (m-MIMO) jest przykładem technologii, która dzięki systemowi sterowania ukierunkowaniem wiązek umożliwia przestrzenną separację łączy użytkowników sieci, korzystających z tej samej stacji dostępowej [5]. Praktyczne wykorzystanie tej technologii wymaga jednakże rozwiązania dodatkowego problemu związanego z koniecznością oceny ograniczeń dotyczących wzajemnego ukierunkowania poszczególnych wiązek systemu antenowego. Wynika to z faktu występowania interferencji międzywiązkowych, które degradują jakość transmisji informacji w poszczególnych łączach. Dotychczasowa procedura symulacyjna zgodna ze standardem 3GPP (*3rd Generation Partnership Project*) [1] znajduje powszechne zastosowanie do oceny wymaganej separacji kątowej wiązek, która zapewnia dopuszczalny poziom interferencji międzykanałowych. Jednakże duża złożoność zarówno strukturalna jak i numeryczna tej procedury oraz brak możliwości odejścia od ściśle zdeterminowanych scenariuszy badawczych uniemożliwia ocenę poziomu interferencji międzywiązkowych dla dowolnych warunków środowiskowych. Złożoność analitycznego opisu charakterystyk antenowych i zmienność ich parametrów, która występuje wraz ze zmianą ich ukierunkowania jest dodatkowym czynnikiem, który utrudnia ocenę wymaganej kątowej separacji wiązek. Stąd też, zapewnienie uproszczenia procedury oceny zakłóceń międzywiązkowych ma duże znaczenie dla procesu przestrzennego planowania i zarządzania w bezprzewodowych sieciach, które bazują na technologii m-MIMO. I tak na przykład, wykorzystanie algorytmu bazującego na wieloelipsoidalnym modelu propagacji MPM daje możliwość adaptacji badań symulacyjnych do dowolnych warunków środowiskowych. Zbieżność wyników oceny poziomu interferencji międzywiązkowych uzyskanych na bazie procedury 3GPP i z wykorzystaniem MPM (*multi-ellipsoidal propagation model*) pokazana jest w [3, 4].

Ten artykuł również poświęcony jest metodyce oceny zakłócającego oddziaływania sygnałów pochodzących od wiązek sąsiadujących z wiązką użyteczną. Jednakże w tym przypadku, główny problem jest skoncentrowany na ocenie możliwości uproszczenia złożonych charakterystyk rzeczywistych systemów antenowych za pomocą modeli standardowych takich jak: charakterystyka gaussowska i jej modyfikacja [7]. W przypadku rzeczywistych anten ze sterowanymi wiązkami, wraz ze zmianą ich ukierunkowania zmianie ulegają zarówno parametry wiązek jak i ich kształty. Ten fakt w znacznym stopniu potęguje złożoność numeryczną prowadzenia badań symulacyjnych i analityczną ocenę poziomu interferencji pochodzących od wspólnokanałowych wiązek. W tym artykule pokazano, że odpowiednia modyfikacja standardowych modeli wiązek względem rzeczywistych charakterystyk antenowych zapewnia uzyskanie zbieżnych wyników w ocenie poziomu interferencji. Ta modyfikacja polega na uwzględnieniu jedynie zmiany poziomu zysku modelu wiązki antenowej proporcjonalnej do zmiany tego parametru dla wiązki rzeczywistej. Zmiana ukierunkowania wiązki z zachowaniem jej kształtu w istotnym stopniu upraszcza procedurę prowadzenia badań symulacyjnych i daje możliwość oceny wpływu poszczególnych jej elementów na poziom indukowanych w antenie odbiorczej

interferencji. Prezentowana w artykule procedura wykorzystuje modyfikację standardowych modeli charakterystyk wiązek antenowych do analizy poziomu zakłóceń interferencyjnych występujących w łączach radiowych z wielowiązkowymi systemami antenowymi. Wykorzystanie uproszczonych charakterystyk do oceny skutków oddziaływania poszczególnych wiązek antenowych na odbierany sygnał użyteczny decyduje o oryginalności opracowanego rozwiązania.

2. CHARAKTERYSTYKA ANTENY WIELOWIĄZKOWEJ 3GPP

Ocena poziomu interferencji między wiązkowych w systemach bezprzewodowych 5G bazuje na modelu charakterystyki promieniowania mocy anteny, który rekomendowany jest przez 3GPP. Ten model opisuje właściwości transmisyjne szyku antenowego w postaci uporządkowanego zbioru elementów promieniujących, który tworzy N_H szeregów i N_V kolumn. W tym przypadku, analityczna postać charakterystyki promieniowania mocy szyku antenowego jest następująca [2]:

$$|g_0(\theta,\phi)|^2 = A(\theta,\phi) + 10log\left(\frac{1}{K}\sum_{k=1}^{K}\left|\sum_{m=1}^{N_H}\sum_{n=1}^{N_V} w_{k,n,m}(\theta,\phi)\cdot v_{n,m}(\theta,\phi)\right|^2\right) \quad (1)$$

gdzie: $A(\theta,\phi)$ jest charakterystyką promieniowania mocy elementu szyku antenowego, K jest liczbą wiązek szyku antenowego, $v_{n,m}$ są współczynnikami szyku, które związane są z położeniem elementów promieniujących a $w_{k,n,m}$ są współczynnikami wagowymi poszczególnych współczynników szyku, które zależą od ukierunkowania szyk antenowego [2].

Charakterystyka promieniowania każdego z elementów szyku antenowego opisana jest zależnością:

$$A(\theta,\phi) = G_{max} - \min\left(-\left(A_H(\phi) + A_V(\theta)\right), G_0\right) \quad (2)$$

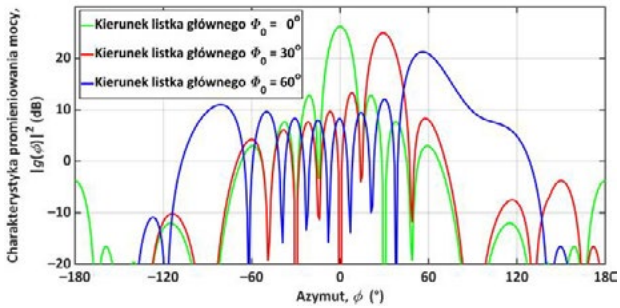
gdzie $G_{max} = 7{,}5$ dBi reprezentuje zysk elementu promieniującego i jest wartością stałą, $A_H(\phi)$ i $A_V(\theta)$ są charakterystykami elementu promieniującego odpowiednio w płaszczyźnie azymutu i elewacji a G_0 jest wartością stałą ($G_0 = 30$ dB). Natomiast charakterystyki $A_H(\phi)$ i $A_V(\theta)$ definiowane są w mierze logarytmicznej następująco:

$$A_H(\phi) = -min\left(12\left(\frac{\phi}{\phi_{3dB}}\right)^2, G_0\right),$$
$$A_V(\theta) = -min\left(12\left(\frac{\theta-90°}{\theta_{3dB}}\right)^2, G_0\right) \quad (3)$$

gdzie ϕ_{3dB} i θ_{3dB} reprezentują kątową rozwartość charakterystyk na poziomie -3 dB odpowiednio w płaszczyźnie azymutu i elewacji tj. $A_H(\phi_{3dB}/2) = -3$ dB i $A_H\left((\theta_{3dB}/2) + 90°\right) = -3$ dB. W prezentowanej analizie, wartości $\phi_{3dB} = 80°$ i $\theta_{3dB} = 65°$ są przyjęte na podstawie rekomendacji [2].

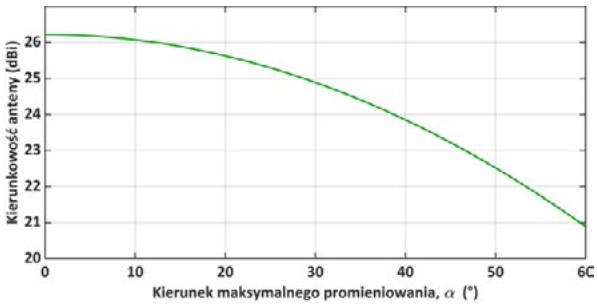
Analiza prezentowana w tym artykule bazuje na szyku antenowym o wymiarach 12×8 elementów. W tym przypadku rozwartość listka głównego wynosi 12,6° i 6° odpowiednio w płaszczyźnie elewacji i azymutu.

Niestety, charakterystyka promieniowania szyku antenowego, która jest generowana na podstawie zależności (1), ulega deformacjom wraz ze zmianą kierunku listka głównego. Na rys. 1 przedstawione są w płaszczyźnie azymutu deformacje charakterystyki, które występują dla wybranych kierunków listka głównego tj. dla kątów $\Phi_0 = 0°$, $\Phi_0 = 30°$ i $\Phi_0 = 60°$.



Rys. 1. Deformacje charakterystyki promieniowania wraz ze zmianą kierunku listka głównego

Można zauważyć, że wraz ze zmianami kierunku listka głównego zmianie ulega również wartość kierunkowości szyku antenowego. Dla analizowanego zbioru parametrów konstrukcyjnych, zmiany wartości kierunkowości szyku przedstawione są na rys. 2.



Rys. 2. Zmiany wartości kierunkowości anteny wraz w funkcji kierunku maksymalnego promieniowania

Zmiana kształtu charakterystyki promieniowania i parametrów transmisyjnych szyku antenowego, która występuje wraz ze zmianą ukierunkowania listka głównego, w znacznym stopniu utrudnia ocenę międzywiązkowych zakłóceń. To wynika z faktu, że każdorazowo dla zmieniających się położeń obiektów, proces generacji charakterystyki szyku antenowego musi być zrealizowany. W tym bowiem przypadku, kątowe przesunięcie charakterystyki referencyjnej nie zapewnia dokładnego odwzorowania kątowego rozkładu mocy zarówno sygnału użytecznego jak i zakłócającego, który pochodzi od pozostałych wiązek abonenckich szyku antenowego.

Z rys. 2 wynika, że aby odwzorować właściwości charakterystyk rzeczywistych anteny należy w przypadku prostych modeli charakterystyk antenowych wprowadzić modyfikację wartości kierunkowości wraz ze zmianą kierunku maksymalnego promieniowania.

3. PROSTE MODELE CHARAKTERYSTYK ANTENOWYCH

Do analitycznego opisu prostych modeli charakterystyk antenowych wykorzystywane są funkcje elementarne. W [7] do podstawowych modeli wykorzystywanych do aproksymacji rzeczywistych charakterystyk anten autorzy zaliczyli unormowaną charakterystykę gaussowską:

$$g(\phi,\theta) = exp\left(-2ln2\left(\frac{\phi}{HPBW_\phi}\right)^2\right) exp\left(-2ln2\left(\frac{\theta}{HPBW_\theta}\right)^2\right) \quad (4)$$

gdzie: $HPBW_\phi$, $HPBW_\theta$ (*half power beam width*) oznaczają szerokość listka głównego odpowiednio w płaszczyźnie azymutu i elewacji.

4. ZMODYFIKOWANY MODEL GAUSSA

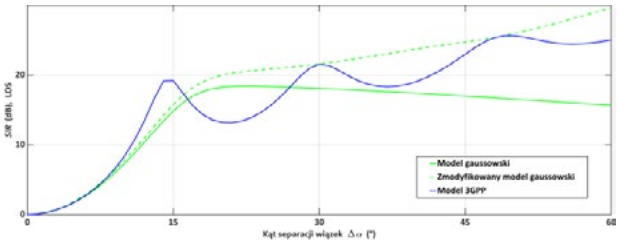
Modyfikacja modelu gaussowskiego polega na wprowadzeniu współczynnika korekcyjnego pokazanego na rys. 2. W tym przypadku, dla analizowanego kierunku promieniowania anteny, maksymalna wartość charakterystyki opisanej zależnością (4) jest równa wartości kierunkowości systemu antenowego charakterystyki rzeczywistej dla danego α określonej na wykresie (rys. 2). Poza tym, kształt charakterystyk modeli zmodyfikowanych w stosunku do prostych nie ulega zmianie.

Wprowadzenie modyfikacji pozwala na zmniejszenie wartości błędu średniokwadratowego (RMSE) średnio o 5 dB niezależnie od kierunku promieniowania anteny.

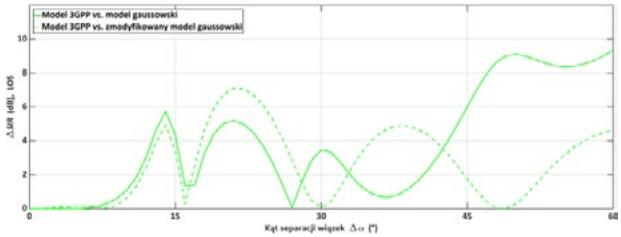
5. BADANIA SYMULACYJNE

W przeprowadzonych badaniach przyjęto scenariusz symulacyjnych dla łącza w dół (*downlink*) oraz założenia analogiczne jak w [3]. Jako podstawowe miary oceny interferencji międzywiązkowych wykorzystano stosunek sygnału do interferencji *SIR* [3, 4] oraz błąd tej miary ΔSIR wyznaczany jako różnice wartości *SIR* określonych dla analizowanego modelu (prostego lub zmodyfikowanego) oraz modelu 3GPP dla technologii m-MIMO jako modelu referencyjnego (oznaczonego kolorem niebiskim na wykresach). Parametry te były analizowane dla kąta separacji wiązek antenowych $\Delta\alpha$ w przedziale od 0 do 60°. Do wyznaczania *SIR* wykorzystano charakterystyki kątowego rozproszenia mocy, które były otrzymane za pomocą MPM.

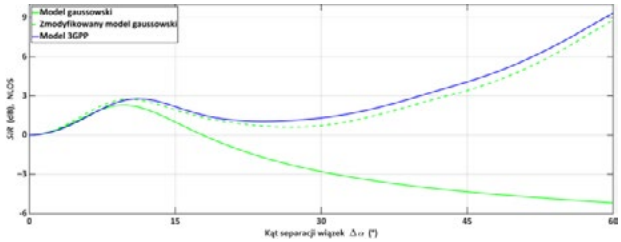
Otrzymane wyniki badań symulacyjnych przedstawiono na rys. 3–6. Rysunki 3 i 5 ilustrują *SIR* w funkcji $\Delta\alpha$ odpowiednio dla warunków LOS (*line-of-sight*) oraz NLOS *(non-LOS*). Natomiast, błąd ΔSIR wyznaczany względem wyników uzyskanych dla charakterystyki 3GPP [3] zilustrowano na rys. 4 i 6 odpowiednio dla LOS i NLOS.



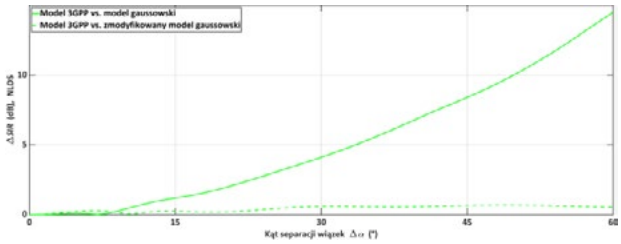
Rys. 3. SIR w funkcji kąta separacji wiązek anteny $\Delta\alpha$ dla warunków LOS



Rys. 4. Błąd SIR w funkcji kąta separacji wiązek anteny $\Delta\alpha$ dla warunków LOS



Rys. 5. SIR w funkcji kąta separacji wiązek anteny $\Delta\alpha$ dla warunków NLOS



Rys. 6. Błąd SIR w funkcji kąta separacji wiązek anteny $\Delta\alpha$ dla warunków NLOS

Uzyskane wyniki pokazują, że wykorzystanie zmodyfikowanego modelu gaussowskiego pozwala na dokładniejsze odzwierciedlenie *SIR* w funkcji $\Delta\alpha$ niż w przypadku modelu prostego. Jest to szczególnie widoczne dla warunków NLOS, gdzie dla modeli zmodyfikowanych uzyskano $\Delta SIR < 1$ dB w całym analizowanym przedziale zmienności $\Delta\alpha$. Lepsze dopasowanie wyników dla warunków NLOS wynika z faktu mniejszego wpływu listków bocznych charakterystyk antenowych na uzyskiwane charakterystyki kątowego rozproszenia mocy i tym samym wartości *SIR* [3].

6. PODSUMOWANIE

Celem artykułu była prezentacja sposobu wykorzystania prostego modelu charakterystyk anten oraz jego modyfikacji na potrzeby oceny interferencji międzywiązkowych w systemach m-MIMO. Do oceny interferencji wykorzystano miarę *SIR*, natomiast do oceny poprawności odzwierciedlenia tej miary w stosunku do wyznaczanej dla charakterystyki rzeczywistej 3GPP użyto miarę ΔSIR. Zaproponowana przez autorów modyfikacja modelu Gaussa pokazała, że zdecydowanie lepiej odzwierciedla ona *SIR* w warunkach NLOS niż model niezmodyfikowany.

Autorzy planują przedstawić rozszerzenie niniejszego referatu w [8].

LITERATURA

[1] 3GPP. 2022. “5G. Study on channel model for frequencies from 0.5 to 100 GHz”, TR 38.901 V17.0.0.

[2] 3GPP. 2019. “Evolved Universal Terrestrial Radio Access (E-UTRA) and Universal Terrestrial Radio Access (UTRA; Radio Frequency (RF) requirement back-ground for Active Antenna System (AAS) Base Station (BS)”, TR 37.842 V13.3.0.
[3] Bechta Kamil, Jan M. Kelner, Cezary Ziółkowski, Leszek Nowosielski. 2021. “Inter-beam co-channel downlink and uplink interference for 5G New Radio in mm-wave bands”. *Sensors*, 21 (3) : e793.
[4] Bechta Kamil, Cezary Ziółkowski, Jan M. Kelner, Leszek Nowosielski. 2021. “Modeling of downlink interference in massive MIMO 5G macro-cell”. *Sensors*, 21 (2) : e597.
[5] Eldowek Basim Mohammed, Nariman Abd El-Salam Bauomy, El-Sayed M. El-Rabaie, Saied M. Abd El-Atty, Fathi E. Abd El-Samie. 2021. “A survey of 5G millimeter wave, massive multiple-input multiple-output, and vehicle-to-vehicle channel measurements and models”. *International Journal of Communication Systems*, 34 (16) : e4830.
[6] Hemadeh Ibrahim A., Katla Satyanarayana, Mohammed El-Hajjar, Lajos Hanzo. 2018. “Millimeter-wave communications: Physical channel models, design considerations, antenna constructions, and link-budget”. *IEEE Communications Surveys Tutorials*, 20 (2) : 870–913.
[7] Vaughan, R.; Bach Andersen, J. *Channels, Propagation and Antennas for Mobile Communications*; IET Electromagnetic Waves Series; Institution of Engineering and Technology: London, UK, 2003; ISBN 978-0-86341-254-7.
[8] Wojtuń, J.; Ziółkowski, C.; Kelner, J.M. 2023. “Inter-beam Interference Assessment Based on Modification Antenna Pattern Models”. Sensors. (w recenzji).

ANALIZA PORÓWNWAWCZA METOD ESTYMACJI MACIERZY KOWARIANCJI ZAKŁÓCEŃ W PRZESTRZENNO-CZASOWYM ADAPTACYJNYM PRZETWARZANIU SYGNAŁÓW

CPMPARATIVE ANALYSIS OF METHODS FOR ESTIMATING THE CLUTTER COVARIANCE MATRIX IN SPACE-TIME ADAPTIVE PROCESSING

Błażej Ślesicki[1]; Anna Ślesicka[2]; Adam Kawalec[3]; Małgorzata Żmigrodzka[4]

[1] Wydział Lotnictwa, Katedra Awioniki i Systemów Sterowania, Lotnicza Akademia Wojskowa, Dęblin, b.slesicki@law.mil.pl
[2] Instytut Nawigacji, Lotnicza Akademia Wojskowa, Dęblin, a.slesicka@law.mil.pl
[3] Wydział Mechatroniki, Uzbrojenia i Lotnictwa, Instytut Techniki Rakietowej i Mechatroniki, Zakład Konstrukcji Rakietowych, Wojskowa Akademia Techniczna, Warszawa, adam.kawalec@wat.edu.pl
[4] Wydział Bezpieczeństwa Lotniczego, Katedra Bezpieczeństwa Transportu Lotniczego, Lotnicza Akademia Wojskowa, Dęblin, m.zmigrodzka@law.mil.pl

DOI: 10.15199/59.2023.4.66

Streszczenie: **Artykuł zawiera analizę porównawczą statystycznych i niestatystycznych metod estymacji macierz kowariancji zakłóceń w przestrzenno-czasowym adaptacyjnym przetwarzaniu (ang. Space-Time Adaptive Processing STAP) sygnału radarowego dla modelu radaru MIMO (ang. Multiple Input Multiple Output). Zaprezentowano istotę, przebieg algorytmu STAP oraz właściwości najnowszych typów metod estymacji macierzy kowariancji zakłóceń. Autorzy przytoczyli w artykule również swój wkład w rozwój tej technologii.**

Abstract: **The article presents a comparative analysis of statistical and non-statistical methods of estimating the clutter covariance matrix in Space-Time Adaptive Processing by using Multiple Input Multiple Output radar geometry. In addition, the properties and stages of the STAP algorithm are presented, as well as the main features of the latest methods for estimating the clutter covariance matrix. Moreover, the authors referenced their contributions to the development of this technology in the article.**

Słowa kluczowe: **przestrzenno-czasowe adaptacyjne przetwarzanie sygnałów, STAP, radar, MIMO.**

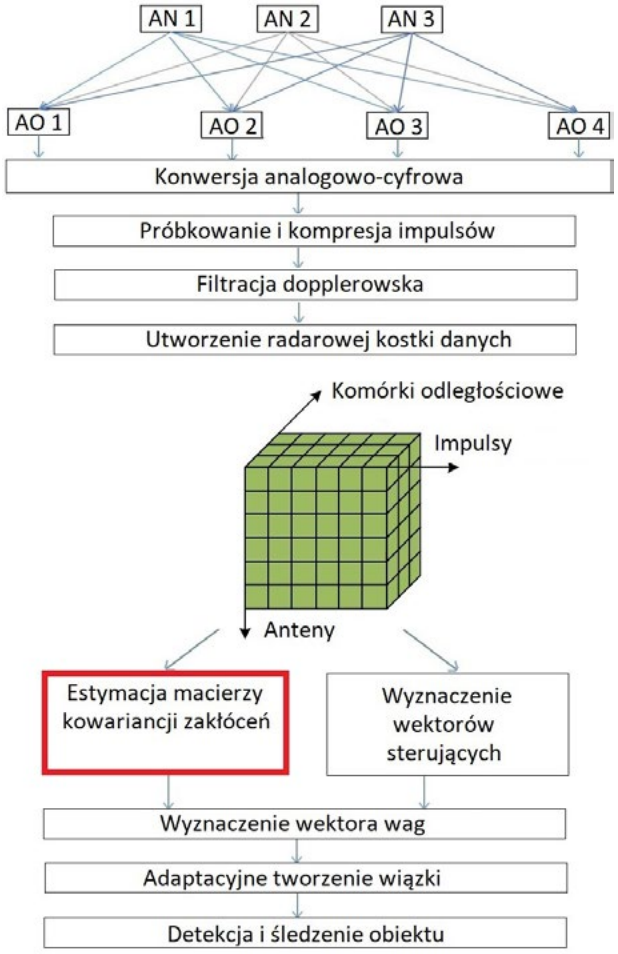
Keywords: **space-time adaptive processing, STAP, signal processing, radar, MIMO.**

1. WSTĘP

Przestrzenno-czasowe adaptacyjne przetwarzanie (STAP) sygnałów radarowych jest jedną z najpopularniejszych technik stosowaną do wykrywania za pomocą radaru umieszczonego na platformie latającej wolno poruszających się obiektów naziemnych na tle silnych zakłóceń pasywnych oraz aktywnych [1-4]. Technologia STAP jest skojarzeniem dwóch technik: technik tłumienia ech stałych, które pracują tylko w dziedzinie czasu oraz techniki związanej z antenami adaptacyjnymi pracującej w dziedzinie przestrzennej w celu eliminacji zakłóceń mających charakter kierunkowy [5-9].

Na rysunku 1 przedstawiono kolejne etapy przetwarzania STAP. Kluczowym etapem w algorytmie STAP jest poprawne wyznaczenie macierzy kowariancji zakłóceń. W literaturze przedmiotu można zauważyć wyraźny podział metod estymacji macierzy kowariancji zakłóceń na metody statystyczne oraz niestatystyczne [5].

Celem artykułu jest zaprezentowanie właściwości obu typów metod estymacji macierzy kowariancji zakłóceń w przetwarzaniu STAP dla radaru MIMO.

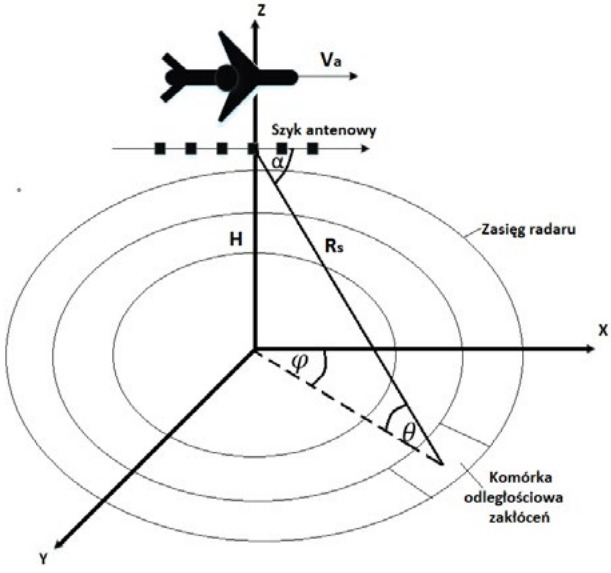


Rys. 1. Schemat przetwarzania STAP (AN – antena nadawcza, AO – antena odbiorcza) [1-4].

2. MODEL SYSTEMU RADAROWEGO

2.1. Model radaru

Rozważono model radaru MIMO zamontowany na platformie latającej. Radar MIMO składa się z *M*-elementowego szyku anten nadawczych oraz *N*-elementowego szyku anten odbiorczych. Każda element liniowego szyku antenowego jest izotropowy. Odległość pomiędzy antenami nadawczymi oznaczono jako d_t, zaś odległość pomiędzy antenami odbiorczymi jako d_r. Model zakłada transmisję impulsów z multipleksacją w dziedzinie czasu (ang. time division multiplexing TDM). Wykorzystanie tak skonstruowanego modelu radaru MIMO zwiększa możliwości detekcji obiektu [10]. Obiekt oznaczony na rysunku jako *P* znajduje się względem radaru w odległości R_s, kąt elewacji pomiędzy nimi wynosi θ zaś azymutu φ. Samolot lecący na wysokości *H*, porusza się wzdłuż linii prostej ze stałą prędkością V_a, a umieszczony na nim radar MIMO opromieniowuję obiekt *P* pod kątem α [9-10].



Rys. 2. Model geometrii radaru [5].

2.2. Model sygnału

Radar pracuje na długości fali λ, nadając ciąg *C* koherentnych impulsów o częstotliwości powtarzania impulsów równej f_r. W każdym z *N*-elementowego szyku anten odbiorczych znajduje się *M* filtrów dopasowanych do odbioru ortogonalnych sygnałów. Stąd, całkowita liczba odebranych sygnałów jest równa iloczynowi *MN*. Macierz odebranych danych wyrażona jest jako [5-8]:

$$\boldsymbol{X} = \boldsymbol{AS} + \boldsymbol{Z} \tag{1}$$

$$\boldsymbol{A} = \boldsymbol{A_r} \odot \boldsymbol{A_t} \tag{2}$$

gdzie $\boldsymbol{A}$ oznacza wirtualny szyk antenowy, $\boldsymbol{A_t}$ oznacza wektor sterujący szyku nadawczego, zaś $\boldsymbol{A_r}$ oznacza wektor sterujący szyku odbiorczego [6]:

$$\boldsymbol{A}_r = \left[\boldsymbol{a}_r(\theta_1), \boldsymbol{a}_r(\theta_2), \dots, \boldsymbol{a}_r(\theta_p)\right] \tag{3}$$

$$\boldsymbol{a}_r(\theta_P) = \left[1, e^{-j2\pi\frac{d_r \sin\theta_P}{\lambda}}, \dots, e^{-j2\pi(N-1)\frac{d_r \sin\theta_P}{\lambda}}\right]^T \tag{4}$$

$$\boldsymbol{A}_t = \left[\boldsymbol{a}_t(\theta_1), \boldsymbol{a}_t(\theta_2), \dots, \boldsymbol{a}_t(\theta_p)\right] \tag{5}$$

$$\boldsymbol{a}_t(\theta_P) = \left[1, e^{-j2\pi\frac{d_t \sin\theta_P}{\lambda}}, \dots, e^{-j2\pi(M-1)\frac{d_t \sin\theta_P}{\lambda}}\right]^T \tag{6}$$

$\boldsymbol{S}$ oznacza macierz odebranych danych od obiektu, która również zawiera zakłócenia pasywne oraz aktywne. $\boldsymbol{X}$ jest macierzą danych dla danego wirtualnego szyku antenowego. $\boldsymbol{Z}$ reprezentuje macierz białego szumu Gaussowskiego. Stąd, biorąc pod uwagę powyższe [6]:

$$\boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \dots, \boldsymbol{s}_C] \tag{7}$$

$$\boldsymbol{X} = \left[\boldsymbol{x}_{1,1}, \dots, \boldsymbol{x}_{M,N}\right] \tag{8}$$

$$\boldsymbol{x}_{m,n} = \left[x_{m,n}(1), x_{m,n}(2), \dots, x_{m,n}(C)\right] \tag{9}$$

$$\boldsymbol{Z} = \left[\boldsymbol{z}_{1,1}, \dots, \boldsymbol{z}_{M,N}\right] \tag{10}$$

$$\boldsymbol{z}_{m,n} = \left[z_{m,n}(1), z_{m,n}(2), \dots, z_{m,n}(C)\right] \tag{11}$$

Macierz kowariancji zakłóceń wyrażona jest jako [6]:

$$\boldsymbol{R}_x = E\{\boldsymbol{X}\boldsymbol{X}^H\} \tag{12}$$

$$\boldsymbol{R}_x = \boldsymbol{A}\boldsymbol{R}_s\boldsymbol{A}^H + \boldsymbol{R}_N \tag{13}$$

$$\boldsymbol{R}_s = E\{\boldsymbol{S}\boldsymbol{S}^H\} \tag{14}$$

gdzie $\boldsymbol{R_S}$ oznacza macierz kowariancji odebranego sygnału, zaś $\boldsymbol{R}_N$ oznacza macierz kowariancji szumu.

Procesor STAP jest filtrem liniowym, więc ostatecznie jego zadaniem jest usunięcie zakłóceń i wykrycie celu. Zależność opisująca te czynność podana jest jako [5-8]:

$$\boldsymbol{Y} = \boldsymbol{w}^H\boldsymbol{X} \tag{15}$$

gdzie $\boldsymbol{w}$ oznacza wektor wag dany zależnością:

$$\boldsymbol{w} = \boldsymbol{R}_x^{-1}\boldsymbol{A} \tag{16}$$

Ostatnim etapem przetwarzania STAP jest zdefiniowanie oraz zweryfikowanie hipotez o obecności H_0 lub braku obecności pożądanego obiektu H_1. Statystyka testowa na obecność celu dana jest jako [6-8]:

$$\eta = \frac{|\boldsymbol{A}^H R_x^{-1}\boldsymbol{X}|}{\boldsymbol{A}^H R_x^{-1}\boldsymbol{A}} \tag{17}$$

Ostatecznie, otrzymaną wartość η porównuje się z przyjętą wartością progu decyzyjnego γ. Sytuacja gdy $\eta < \gamma$ oznacza brak obiektu, w przeciwnym razie obiekt został wykryty.

3. METODY ESTYMACJI MACIERZY KOWARIANCJI ZAKŁÓCEŃ

3.1. Metody statystyczne

Klasyczną, statystyczną metodą estymacji macierzy kowariancji zakłóceń jest metoda SMI STAP (ang. Sample Matrix Inversion STAP). Metoda SMI została szczegółowo opisana w [4]. Metoda SMI STAP stanowi najczęściej metodę odniesienia do nowo proponowanych statystycznych metod estymacji macierzy kowariancji zakłóceń w STAP [7-9].

Ponadto, wśród metod statystycznych można wyróżnić metody polegające na zamierzonym doborze komórek treningowych [11-13]. W artykule [12] opisano metodę

wyboru komórek treningowych TSS (ang. Training Sample Selection) polegającą na przyjęciu tylko tych komórek treningowych, dla których macierze kowariancji są podobne w określony sposób do macierzy kowariancji komórki testowanej na obecność obiektu.

Inną stosowaną metodą wyboru komórek treningowych w celu wyznaczenia macierzy kowariancji zakłóceń jest metoda GIP (ang. Generalized Inner Product). Szczegółowy opis tej metody można znaleźć w [14].

Z czasem powstały metody statystyczne bazujące na wcześniej wspomnianych, takie jak PSWF-GIP (ang. Prolate Spheroidal Wave Functions GIP) [15] czy CSMS-GIP (ang. Cross-Spectral Metric Smoothing GIP) [16], które również na podstawie pewnej statystyki oraz wcześniej przekształconej macierzy kowariancji zakłóceń całego zbioru komórek treningowych, dokonują wyboru właściwych komórek treningowych .

3.2. Autorska metoda statystyczna STAP

Autorzy niniejszego artykułu biorąc pod uwagę aktualną analizę literatury, zauważyli, że obecnie na świecie skupiono uwagę na badaniu niestatystycznych metod estymacji macierzy kowariancji zakłóceń w środowiskach heterogenicznych. Stąd, w artykule [6] zaproponowano nową, niestatystyczną metodę estymacji macierzy kowariancji zakłóceń. Nowa metoda oparta została na metodzie SMI, z tym, że zamiast rozkładu QR w ciele metody wykorzystano rozkład LU z częściowym podstawieniem elementów macierzy.

3.3. Metody niestatystyczne

Do metod niestatystycznych możemy zaliczyć metodę KA-STAP (ang. Knowledge-Aided STAP) bazującą na wcześniej znanych parametrach wykorzystywanego radaru bądź środowiska oraz metodę bezpośredniej dziedziny danych D3 STAP (ang. Direct Data Domain STAP).

W literaturze przedmiotu stosuje się podział metod KA-STAP na metody bezpośredniego oraz pośredniego wykorzystania zdobytej wcześniej wiedzy. Dane do wyznaczenia macierzy kowariancji zakłóceń pozyskiwane są z wcześniejszych sondowań radaru bądź z analizy mapy skanowanego terenu.

Jednym z bardziej znanych algorytmów KA-STAP jest algorytm CC-KA-STAP (ang. Convex Combination KA-STAP) [17]. Algorytm na samym początku przeszukuje odebrane dane i stara się wykryć jednoznaczne, „silne" obiekty. Algorytm ten można przyporządkować do algorytmów KA-STAP z bezpośrednim wykorzystaniem danych.

Algorytmy D3-STAP polegają na wykorzystywaniu pojedynczego przekroju radarowej kostki danych. Eliminuje się w ten sposób uwzględnianie sąsiednich komórek odległościowych będących komórkami treningowymi. Dwoma typowymi algorytmami D3 STAP jest algorytm D3 LS (ang. direct data domain least square) czyli algorytm D3 w sensie najmniejszych kwadratów [18] oraz D3 SR (ang. direct data domain sparse recovery) czyli algorytm D3 wykorzystujący rzadkość macierzy zakłóceń [19].

W ostatnich kilku latach cały szereg badań skoncentrowany jest w głównej mierze na połączeniu metod D3 STAP wraz z algorytmami rzadkiego odzyskiwania, o czym świadczą publikacje [1,20].

3.4. Autorska metody niestatystyczna STAP

W artykule [3] udowodniono, że zastosowanie modelu geometrii radaru MIMO oraz algorytmu dopasowania kroczącego do estymacji macierzy kowariancji zakłóceń w technice STAP, pozwala na wykrycie obiektu na tle zakłóceń niejednorodnych. W efekcie czego powstała nowa niestatystyczna metoda estymacji macierzy kowariancji zakłóceń OMP-STAP [4].

4. SYMULACJE

W celu porównania wydajności zaprezentowanych metod wykonano symulacje w środowisku Matlab dla parametrów zawartych w tabeli 1.

Tab. 1. Parametry przyjęte do symulacji

Parametr	*Wartość*	*Jednostka*
Częstotliwość radaru	10	[GHz]
Czas trwania impulsu	33	[μs]
Częst. powtrz. impul.	30	[kHz]
Ilość impulsów	10	[-]
Anteny nadawcze	5	[-]
Anteny odbiorcze	5	[-]
Wysokość	1000	[m]
Prędkość platformy	225	[m/s]

Przedstawiono wydajność procesora liniowego STAP bazującego na konkretnych metodach estymacji macierzy kowariancji zakłóceń. W literaturze przedmiotu parametrem definiującym wydajność procesora STAP nazwano współczynnik poprawy IF (ang. Improvement Factor), który określa iloraz stosunku mocy sygnału do szumu na wyjściu i na wejściu optymalnego procesora.

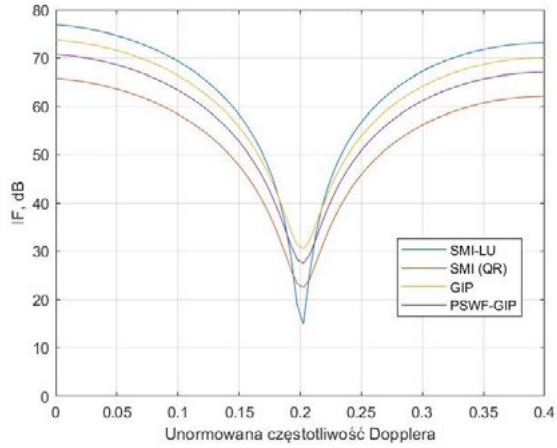
Biorąc pod uwagę wyniki symulacji na obu poniższych rysunkach należy stwierdzić, że proponowane przez autorów rozwiązania cechują się najwyższą wydajnością procesora STAP.

5. WNIOSKI

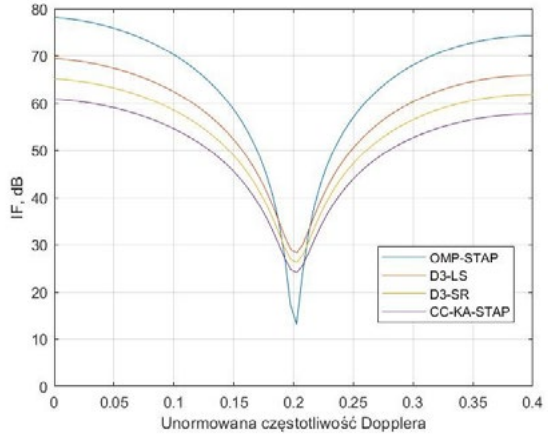
Podsumowując, metody statystyczne są wysoce skuteczne w przypadku występowania zakłóceń o charakterze jednorodnym. Wynika to z założenia, że komórki zakłóceń mają identyczny rozkład prawdopodobieństwa. Jednakże metody statystyczne wymagają dostępu do dużej ilości komórek treningowych, co przeważnie jest trudne do spełnienia.

W ogólności metody niestatystyczne redukują problemy jakie występowały w metodach statystycznych oraz zmniejszają stopień złożoności obliczeniowej. Ponadto, poprzez wykorzystanie małej ilości komórek treningowych lub w niektórych przypadkach jednej, a także poprzez wykorzystanie rzadkości zakłóceń w dziedzinie

przestrzenno-czasowej, możliwe jest uzyskanie wysokiej rozdzielczości przestrzenno-czasowego widma zakłóceń.



Rys. 3.Wydajność procesora STAP dla metod statystycznych..



Rys. 4. Wydajność procesora STAP dla metod niestatystycznych.

LITERATURA

[1] Anna Ślesicka, Adam Kawalec. 2020. „An Application of the Orthogonal Matching Pursuit Algorithm in Space-Time Adaptive Processing". *Sensors*, 20 (12) : 3468.

[2] Anna Ślesicka, Adam Kawalec. 2020. „ Analiza porównawcza algorytmów adaptacyjnego przetwarzania przestrzenno-czasowego w zastosowaniach radiolokacyjnych". *Biuletyn Wojskowej Akademii Technicznej*, 69 (2) : 129-147.

[3] Anna Ślesicka, Adam Kawalec. 2020. „ The performance of a linear STAP processor for radar signal processing". *Procedings of 21st International Radar Symposium (IRS),* 154-156.

[4] Anna Ślesicka. 2021. „ *Zastosowanie algorytmu dopasowania kroczącego do wyznaczania macierzy kowariancji zakłóceń w przestrzenno-czasowym adaptacyjnym przetwarzaniu sygnałów*". Rozprawa doktorska, Wojskowa Akademia Techniczna.

[5] Błażej Ślesicki, Anna Ślesicka, Adam Kawalec. 2022. „The Study of the Possibility of Applying Parallel Programming to the Algorithms of Space-Time Adaptive Processing". *Problems of Mechatronics. Armament, Aviation, Safety Engineering,* 13 (3): 27-42

[6] Błażej Ślesicki, Anna Ślesicka, Adam Kawalec. 2023. „ A New Statistical Method for Determining the Clutter Covariance Matrix in Spatial–Temporal Adaptive Processing of a Radar Signal". *Sensors*, 23 (9) : 4280.

[7] Richard Klemm. 2002. „ *Principles of space-time adaptive processing"*. IEE Publishing.

[8] Joseph R. Guerci. 2003. „ *Space-Time Adaptive Processing for Radar*", Artech House Publishers.

[9] James Ward. 1994. „ *Space-time adaptive processing for airborne radar*". Technical Report No. 1015, MIT Lincoln Laboratory.

[10] Jian Li, Petre Stoica. 2009. „ *MIMO Radar Signal Processing*". Wiley: Hoboken.

[11] Yifeng Wu, Tong Wang, Jianxin Wu, Jia Duan. 2015 „ *Training sample selection for space-time adaptive processing in heterogeneous environments*". IEEE Geosci. Remote Sens. Lett. 12 (4), 691–695.

[12] Huiyong Li, Weiwei Bao, Jinfeng Hu, Julan Xie, Ruixin Liu. 2018. „ *A training samples selection method based on system identification for STAP*". Signal Process., 142, 119–124.

[13] Jinfeng Hu, Jianping Li, Huiyong Li, Keze Li, Jing Liang. 2020. „ *A Novel Covariance Matrix Estimation via Cyclic Characteristic for STAP*". IEEE Geosci. Remote Sens. Lett. 17 (11), 1871-1875.

[14] Bo Tang, Jun Tang, Yingining Peng. 2012. „ *Detection of heterogeneous samples based on loaded generalized inner product method*". Digital Signal Process. 22 (4), 605–613.

[15] Xiaopeng Yang, Yongxu Liu, Xiaona Hu, Teng Long. 2013. „*Robust non-homogeneity detection algorithm based on prolate spheroidal wave functions for space-time adaptive processing*". IET Radar, Sonar and Navigation.. 7, 47–54.

[16] Huadong Yuan, Hong Xu, Keqing Duan, Wenchong Xie, Yongliang Wang. 2019. „*Cross-Spectral Metric Smoothing-Based GIP for Space-Time Adaptive Processing*". IEEE Geosci. Remote Sens. Lett. 16 (9), 1388-1392.

[17] Rui Fa, Rodrigo De Lamare, Vitor Nascimento. 2010. „*Knowledge-aided STAP algorithm using convex combination of inverse covariance matrices for heterogeneous clutter*". Proc. IEEE Int. Conf. Acoust. Speech and Signal Process. 2742–2745.

[18] Wonsuk Choi, Tapan. Sarkar, Hong Wang, Eric Mokole. 2006. „*Adaptive processing using real weights based on a direct data domain least squares approach*". IEEE Transactions on Antennas and Propagation. 54 (1), 182–191.

[19] Ke Sun, Huadong Meng, Yongliang Wang, Xiqin Wang. 2011. „*Direct data domain STAP using sparse representation of clutter spectrum*". Signal Processing, 91 (9), 2222–2236.

[20] Wei Zhang, Ruixue An, Ningyu He, Zishu He, Huiyong Li. 2020. „*Reduced dimension STAP based on sparse recovery in heterogeneous clutter environments*". IEEE Trans. on Aerospace and Electronics Systems. 56 (1). 785-795.

ANALIZA PORÓWNAWCZA MODELI PROFILU GĘSTOŚCI ŁADUNKÓW JONOSFERY WYKORZYSTYWANYCH DO PREDYKCJI SYGNAŁÓW W KANAŁACH KF

COMPARATIVE ANALYSIS OF IONOSPHERE CHARGE DENSITY PROFILE MODELS USED FOR SIGNAL PREDICTION IN KF CHANNELS

Bogdan Uljasz[1]; Cezary Ziółkowski[2]

[1] Wojskowa Akademia Techniczna, Warszawa, bogdan.uljasz@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Warszawa, cezary.ziolkowski@wat.edu.pl

DOI: 10.15199/59.2023.4.67

Streszczenie**: Profil gęstości ładunków jonosfery jest podstawą dla modelowania i analizy jakości transmisji sygnałów w łączach KF (fale krótkie) dalekiego zasięgu. Niniejszy artykuł dotyczy oceny dokładności odwzorowania krzywej gęstości ładunków na bazie pięciu wybranych modeli profili. Opisano miarę i metodę analizy porównawczej warunkujące ocenę błędu aproksymacji danych pomiarowych. Prezentowane wyniki są podstawą do wyboru modelu profilu, który zapewnia skuteczne odwzorowanie wpływu warunków propagacyjnych jonosfery na transmitowane sygnały.**

Abstract**: The ionosphere charge density profile is the basis for modeling and analyzing the quality of signal transmission in long-range HF (ang. *High Frequency*) links. This article concerns the assessment of the accuracy of mapping the charge density distribution based on five selected profile models. The measure and method of comparative analysis conditioning the assessment of the measurement data approximation error are described. The presented results are the basis for the selection of a profile model that provides an effective representation of the influence of ionosphere propagation conditions on the transmitted signals.**

Słowa kluczowe**: krzywa gęstości elektronów, propagacja jonosferyczna, modelowanie matematyczne, analiza porównawcza**

Keywords**: electron density curve, ionospheric propagation, mathematical modelling, comparative analysis**

1. WSTĘP

W dobie globalizacji niezbędna jest wymiana danych na bliskie jak i dalekie odległości. Łączność dalekosiężna LRC (ang. *Long Range Communications*) wykorzystuje różne media transmisyjne: przewodowe, satelitarne i jonosferyczne. Obecnie najpowszechniejsze są łącza przewodowe (światłowodowe i miedziane) oraz satelitarne. W trakcie konfliktów międzynarodowych mocno zagrożone są łącza przewodowe, które narażone są na dywersję [6]. Łączność satelitarna bazuje na satelitach umieszczonych na orbitach geostacjonarnej (GEO: ang. *Geosynchronous Equatorial Orbit*) oraz LEO (ang. *Low Earth Orbit*). W dniu napaści Rosji na Ukrainę Internet satelitarny (GEO) świadczony przez usługodawcę w Europie został przerwany. Uszkodzonych zostało wiele terminali naziemnych [7][1]. Spowodowało to m.in. utratę połączenia z turbinami wiatrowymi (np. 5800 turbin wiatrowych w Niemczech). Atak na terminale był przeprowadzony łączami satelitarnymi poprzez funkcję OTA (ang. *Over The Air*).

Łącza jonosferyczne wykorzystywane są przez radiostacje pracujące w paśmie krótkofalowym. Ten sposób łączności jest najtańszy oraz najmniej podatny na ataki (zakłócenia celowe). Łączność jonosferyczna nie wymaga infrastruktury pośredniczącej między korespondentami. Możliwe problemy z łącznością dalekosiężną przewodową i satelitarną zwiększa zainteresowanie w łączności jonosferycznej. Aby zapewnić zadawalające przepływności danych, niezbędne jest opracowywanie technik transmisji szerokopasmowej. Pod koniec XX wieku osiągnięto przepływność 19,2 kbitów/s wykorzystując pasmo do 9 kHz [9]. W roku 2011 zaproponowano modemy o przepływności do 120 kbitów/s wykorzystujące pasmo do 24 kHz [10]. Rok 2017, to propozycja wykorzystania ciągłego pasma do 48 kHz, które pozwala na przesyłanie danych o przepływności do 240 kbitów/s [11]. W roku 2019 zaproponowano modem, który automatycznie wybiera do transmisji wielokanałowej podpasma 3/6 kHz z pasma o szerokości do 200 kHz (przy maksymalnym łącznym paśmie transmisji do 48 kHz) [12]. We wszystkich przypadkach testowania nowych modemów wykorzystywane są obecnie dwa modele kanału: AWGN i ITU-R F.1487 [14]. Modele te nie dają wystarczających informacji o możliwościach modemów szerokopasmowych. Badania jonosferycznych modemów w kanale rzeczywistym jest długotrwałe i kosztowne.

Celem artykułu jest ocena dokładności odwzorowania krzywej gęstości elektronów $N(z)$, która stanowi podstawę do modelowania skutków zjawisk propagacyjnych występujących w jonosferze a tym samym umożliwia prowadzenie badań w zakresie oceny granicznych możliwości transmisji informacji w kanałach jonosferycznych .

2. ROLA PROFILU GĘSTOŚCI ŁADUNKÓW W MODELOWANIU PROPAGACJI FAL

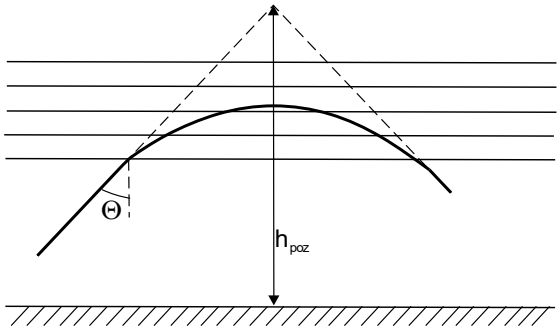
W analizie zjawiska odbicia niezbędna jest ocena wysokości warstwy odbijającej (wysokość pozorna jonosfery). Podstawą do wyznaczenia wysokości pozornej jest profil gęstości ładunku elektronów $N(z)$, który jest ściśle związany ze współczynnikiem refrakcji elektronów n' zależnością [8]:

$$n'(f,z) \cong \left(1 - 81\frac{N(z)}{f^2}\right)^{1/2} \quad (1)$$

Wysokość pozorną liczy się zgonie ze wzorem [8]:

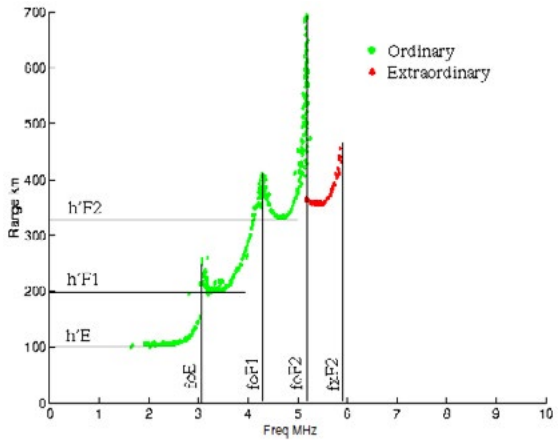
$$h_{poz} = z' = \int_0^{z_0(f)} \frac{\partial}{\partial_f}\left(f \cdot n(f,z)\right) dz = \int_0^{z_0(f)} n'(f,z)dz \quad (2)$$

Biorąc pod uwagę, że odbicie następuje przy wartości współczynnika refrakcji fali *n'=0*, znając częstotliwość sondującą *f* możemy określić gęstość ładunków elektrycznych (elektronów) *N(z)*(2) dla wysokości pozornej odbicia h_{poz} (rys. 1).



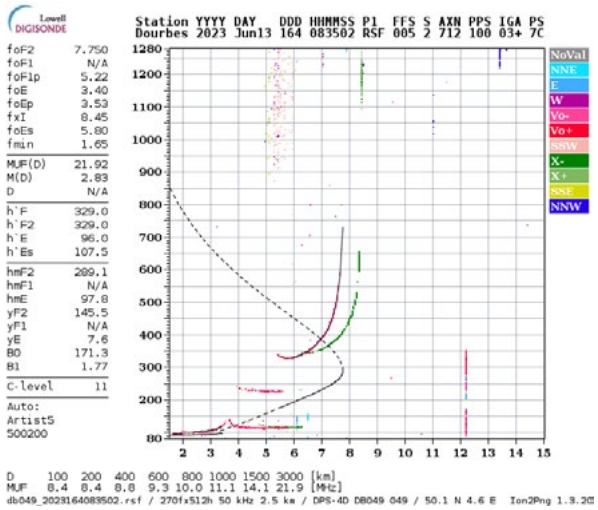
Rys.1. Wysokość pozorna odbicia h_{poz} oraz rzeczywisty przebieg fali radiowej dla kąta wejścia fali Θ [8]

Uzyskanie krzywych gęstości elektronów uzyskuje się przez sondowanie jonosfery sygnałem radiowym. Wynik sondowania pionowego jonosfery przedstawia rysunek 2 (promień zwyczajny - ordinary i nadzwyczajny - extraordinary). Wykorzystując specjalizowane oprogramowanie wyznaczana się krzywą gęstości elektronów oraz parametry charakterystyczne dla poszczególnych warstw, np. wysokość pozorna odbicia dla każdej warstwy (lokalnego maksimum jonizacji): *zF2*, *zF1*, *zE*.



Rys.2. Typowy pionowy jonogram. Wysokość pozorna w funkcji częstotliwości odbiornika. Zaznaczone wartości charakterystyczne [1]

Rysunek 3 przedstawia przykładowy wynik sondowania wykonany przez stację Dourbes w Belgii [4] wraz wyznaczoną krzywą gęstości elektronów i wartościami charakterystycznymi dla poszczególnych warstw. Listę stacji sondujących i ich położenie można pobrać ze strony GIRO [5], ich opis z [14]. Publikowane są wyniki sondowania z odstępem co 15 minut.



Rys. 3. Przykładowy wynik sondowania jonosfery [4]

Jonogramy z różnych stacji sondujących zostały wykorzystane przy tworzeniu bazy danych wejściowych.

3. MODELE PROFILU GĘSTOŚCI ŁADUNKÓW

Modele profili gęstości ładunków przedstawione w [1][2][3][16][17] pozwalają na matematyczne odwzorowanie rzeczywistej krzywej gęstości elektronów na podstawie wartości charakterystycznych jonosfery. Do porównania wybrano pięć modeli, wśród nich najczęściej wykorzystywany model Chapmana. We wszystkich wybranych modelach wykorzystano parametry pobrane z jonogramów (rys. 3):

NmF – gęstość elektronowa dla warstwy F2
NmE – gęstość elektronowa dla warstwy E
zF – wysokość pozorna warstwy F2
zE – wysokość pozorna warstwy E

Poniżej podano wzory dla poszczególnych modeli matematycznych. We wszystkich wzorach zmienna *z* określa wysokość nad powierzchnią Ziemi.

Liniowy model profilu *N(z)* [14]

$$N(z) = a_1 z + b_1 \quad (3)$$

gdzie:

$$a_1 = \frac{N_{mF} - N_{mE}}{z_F - z_E}, \quad (4)$$

$$b_1 = N_{mF} - \frac{N_{mF} - N_{mE}}{z_F - z_E} z_F \quad (5)$$

Wykładniczy model profilu *N(z)* [14]

$$N(z) = a_2 exp b_2 z \quad (6)$$

gdzie:

$$a_2 = N_{mF} exp\left(-\frac{ln\frac{N_{mE}}{N_{mF}}}{z_E - z_F} z_F\right), \quad (7)$$

$$b_2 = \frac{ln\frac{N_{mE}}{N_{mF}}}{z_E - z_F} \quad (8)$$

Paraboliczny model profilu *N(z)* [14]

$$N(z) = a_3 z^2 + b_3 z + c_3 \quad (9)$$

gdzie:

$$a_3 = \frac{N_{mF}-N_{mE}}{(z_F-z_E)^2}, \quad (10)$$

$$b_3 = -2\frac{N_{mE}-N_{mF}}{(z_F-z_E)^2}z_F, \quad (11)$$

$$c_3 = N_{mF} + \frac{N_{mE}-N_{mF}}{(z_F-z_E)^2}z_F^2 \quad (12)$$

Chapmana model profilu $N(z)$ [14]

$$N(z) = N_{mF}exp\frac{1}{2}\left(1 - \frac{z-z_F}{H} - exp\left(-\frac{z-z_F}{H}\right)\right) \quad (13)$$

gdzie H - wysokość minimalna do obliczeń

Model sech² profilu $N(z)$ – hiperboliczny [8]

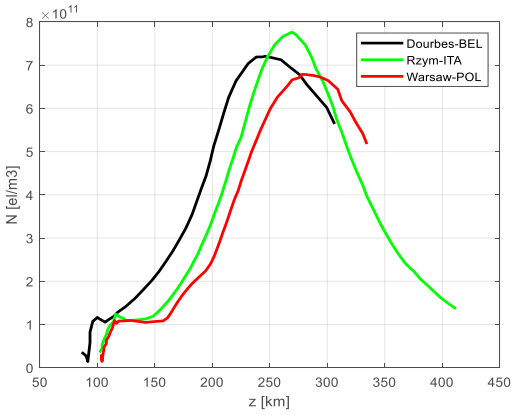
$$N(z) = N_{mF}sech^2\left(\frac{z-z_F}{a}\right) \quad (14)$$

gdzie a - wysokość minimalna do obliczeń

4. ANALIZA PORÓWNAWCZA MODELI

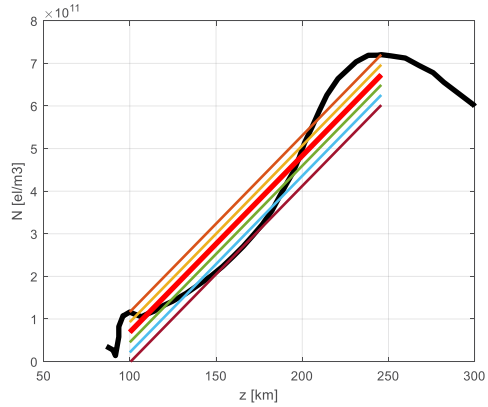
Dla analizy porównawczej modeli istotne było przygotowanie profili referencyjnych na podstawie krzywych z sondowania jonosfery. Pierwszym krokiem przy wyznaczaniu profili referencyjnych było utworzenie cyfrowej bazy danych na bazie wykresów wyników pomiarowych (literaturowych) f(z).

Profile referencyjne przedstawione na rysunku 4 wykonywane były poprzez ucyfrowienie krzywych gęstości elektronów, takich jak przedstawione na rysunku 3.

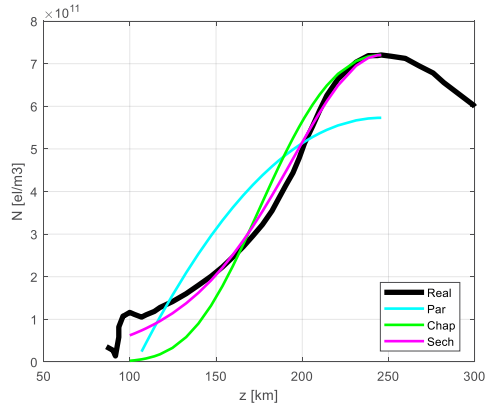


Rys. 4. Przykładowe charakterystyki f=f(h$_{poz}$) wyznaczone na podstawie krzywych gęstości elektronowej – profil referencyjny

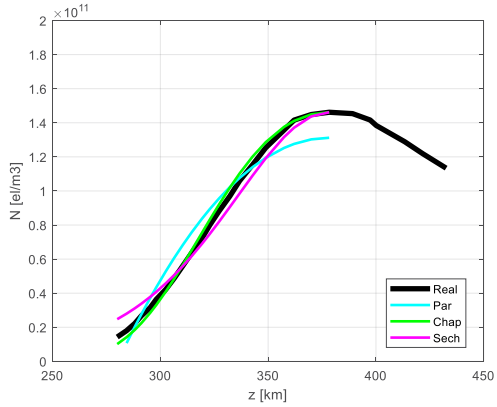
Jako parametry wejściowe modeli profili przyjmuje się wartości f0E, zE, f0F2 i zF2. Na podstawie wartości charakterystycznych określono zakres obliczeniowy do wyznaczenia profili. Następnie zgodnie ze wzorami, podanymi w rozdziale 3 niniejszego artykułu, wyznaczono parametry modeli profili i wykreślono odpowiadające im charakterystyki. Obliczenia wykonano w środowisku Matlab.



Rys. 5. Przykład wyznaczania w krokach najbardziej dopasowanej charakterystyki liniowego modelu profilu N(z)



Rys.6. Wykresy obrazujące położenie profili dla trzech najdokładniejszych modeli profilu N(z) dla danych wejściowych Dourbes Belgia – dzień



Rys.7. Wykresy obrazujące położenie profili dla trzech najdokładniejszych modeli profilu N(z) dla danych wejściowych Dourbes Belgia – noc

Miarę dokładności aproksymacji charakterystyki referencyjnej, wykorzystaną do oceny modeli, przedstawiono we wzorze (15).

$$\Delta_{modelu} = \sqrt{\frac{1}{K}\sum_k\left(\left(N_{p_k} - N_{m_k}\right)^2\right)} \quad (15)$$

gdzie:

K – liczba danych pomiarowych
k – kolejna numer wartości danych pomiarowych N_{pk} *i* N_{mk}
N_{pk} – gęstość jonizacji na podstawie pomiarów jonosfery dla k-tego pomiaru`
N_{mk} – gęstość jonizacji wyliczona dla danego modelu dla k-tego pomiaru

Tab. 1. Ocena dokładności aproksymacji modeli profili (dzień 26.09.2022 godz. 13:30)

Model	*Belgia Dourbes*	*Włochy Rzym*	*Polska Warszawa*
		[10^{+11} el/m^3]	
Paraboliczny	2.1345	3.5540	2.6741
Chapmana	0.6420	1.0401	0.5603
Hiperboliczny	0.4080	0.3196	0.3521

Tab. 2. Ocena dokładności aproksymacji modeli profili (noc 27.09.2022 godz. 00:15)

Model	*Belgia Dourbes*	*Włochy Rzym*	*Polska Warszawa*
		[10^{+11} el/m^3]	
Paraboliczny	0.2643	0.5511	0.1307
Chapmana	0.0282	0.0985	0.0362
Hiperboliczny	0.1054	0.1932	0.0574

Profil o najmniejszej wartości błędu średniokwadratowego uznano jako najbardziej dopasowany do profilu referencyjnego. Dla porównania dwóch najlepszych modeli wykorzystano zależność:

$$\Delta_{pm} = \frac{1}{M}\sum_m \left(\Delta_{chap_m} - \Delta_{sech_m}\right) \qquad (16)$$

gdzie:

Δ_{pm} – różnica między błędami odwzorowania dla analizowanych modeli
M – Liczba analizowanych scenariuszy
m – numer kolejnego scenariusza
$\Delta chap$ – błędy odwzorowania krzywej jonizacji jonosfery dla modelu Chapmana
$\Delta sech$ – błędy odwzorowania krzywej jonizacji jonosfery dla modelu Hiperbolicznego

Tab. 3. Porównanie aproksymacji przez modele Chapmana i Hiperbolicznego dla wszystkich danych

Zakres danych wejściowych	Δ_{pm} *[el/m^3]*
Od fmE do fmF2	$0.1074 \cdot 10^{+11}$
Cały zakres	$0.1616 \cdot 10^{+11}$

5. PODSUMOWANIE

Dla dnia najlepsze odwzorowanie uzyskano dla modelu sech2 (hiperboliczny), dla nocy modelu Chapmana. Dla godzin nocnych przyjęto uśrednione wartości f0E i zE,, będące wynikiem długoletnich obserwacji (rys. 7).

Najmniejszy dobowy błąd odwzorowania profilu referencyjnego uzyskano dla sech^2 profilu N(z) - hiperbolicznego. Wykorzystanie tego modelu wymaga empirycznego wyznaczenia jego parametrów dla możliwie jak najszerszego zakresu warunków propagacyjnych występujących w łączach jonosferycznych oraz ich zmienności w czasie.

LITERATURA

[1] Bamford R.. 2000. “The oblique ionospheric sounder”. *Radiocommunications Agency*, Chilton, Didcot, Wielka Brytania
[2] Budden K.G. 1988. “The propagation of radio waves. The theory of radio wave of low power in the ionosphere and magnetosphere”. *Cambridge University Press*, New York, USA.
[3] Budden K.G. 1966. „Waves in the Ionosphere. The mathematical theory of the reflection of radio waves from stratified Ionized Layers”. *Cambridge University Press*, Wielka Brytania
[4] Dourbes, Belgium, URSI code: DB049, Diginosonde, *https://digisonde.oma.be/*
[5] Global Ionosphere Radio Observatory, https://giro.uml.edu/
[6] Golański A. 2023. „Pentagon ostrzega: Rosja może przeciąć oceaniczne kable telekomunikacyjne”, *https://www.dobreprogramy.pl/pentagon-ostrzega-rosja-moze-przeciac-oceaniczne-kable-telekomunikacyjne,6628278887622785a*
[7] Haertle A. 2023. „Tysiące terminali Internetu satelitarnego poważnie uszkodzonych w dniu ataku na Ukrainę”, *https://zaufanatrzeciastrona.pl/post/tysiace-terminali-internetu-satelitarnego-powaznie-uszkodzonych-w-dniu-ataku-na-ukraine/*
[8] Lisiecki W. 1962. „Propagacja fal radiowych”, *Wydawnictwo Komunikacji i Łączności* (WKiŁ), Warszawa
[9] Norma MIL-STD-188-110B, 2000
[10] Norma MIL-STD-188-110C, 2011
[11] Norma MIL-STD-188-110D, 2017
[12] Norma STANAG 4539, 2019
[13] Połowianiuk M. 2023. „Internet satelitarny przestał działać nad Europą pierwszego dnia wojny. Tysiące terminali zostało zniszczonych”, *https://spidersweb.pl/2022/03/internet-satelitarny-przestal-dzialac.html*
[14] Recommendation ITU-R F.1487 (05/2000), “Testing of HF modems with bandwidths of up to about 12 kHz using ionospheric channel simulators”, *International Telecommunication Union*
[15] Reinisch B.W., Galkin I.A. 2011. “Global Ionospheric Radio Observatory (GIRO)”, *Earth Planets and Space*, DOI: 10.5047/eps.2011.03.001
[16] Stankov, S. M., Jakowski, N., Heise, S., Muhtarov, P., Kutiev, I., and Warnant, R., (2003), A new method for reconstruction of the vertical electron density distribution in the upper ionosphere and plasmasphere, J. Geophys. Res., 108, 1164, doi:10.1029/2002JA009570
[17] Stankov, S.M. Evaluation of analytical ionospheric models used in electron density profile reconstruction. Acta Geod. Geoph. Hung 37, 385–401 (2002). https://doi.org/10.1556/AGeod.37.2002.4.2

ANTENA ENDOSKOPU KAPSUŁKOWEGO PRZEZNACZONA DO PRACY W PASMIE MICS 400 MHz

WIRELESS ENDOSCOPE ANTENNA FOR 400 MHz MICS BAND

Paweł Oleksy[1]; Łukasz Januszkiewicz[2];

[1] Politechnika Łódzka, Instytut Elektroniki, Łódź, pawel.oleksy@p.lodz.pl
[2] Politechnika Łódzka, Instytut Elektroniki, Łódź, lukasz.januszkiewicz@p.lodz.pl

DOI: 10.15199/59.2023.4.68

Streszczenie: **Endoskopy kapsułkowe to miniaturowe urządzenia elektroniczne służące do rejestracji obrazu wnętrza przewodu pokarmowego. W artykule przedstawiono projekt anteny endoskopu kapsułkowego pracującej w pasmie przeznaczonym do zastosowań medycznych: MICS 400 MHz. Opracowana antena charakteryzuje się dobrym dopasowaniem impedancyjnym w różnych obszarach przewodu pokarmowego człowieka. Konstrukcja anteny umożliwia dopasowanie jej kształtu do obudowy endoskopu kapsułkowego. Przedstawiono wyniki symulacji oraz pomiarów parametrów prototypowej anteny.**

Abstract: **Capsule endoscopes are miniature electronic devices used to record the image of the inside of the digestive tract. The article presents the design of an antenna designed to work in a capsule endoscope in the band dedicated for medical applications: MICS 400 MHz. The proposed antenna is characterized by good impedance matching in various areas of the human digestive tract. The design of the antenna allows its shape to be adapted to the housing of the capsule endoscope. The results of simulations and measurements of the parameters of the prototype antenna are presented.**

Słowa kluczowe: **Anteny, endoskopy bezprzewodowe, BAN**

Keywords: **Antennas, wireless endoscopes, BAN**

1. WSTĘP

W diagnostyce chorób układu pokarmowego często wykorzystywaną techniką obrazowania jest endoskopia. Umożliwia ona zidentyfikowanie zmian chorobowych powstałych we wnętrzu przewodu pokarmowego przy użyciu miniaturowych kamer. Obecnie badanie endoskopowe może być przeprowadzone z wykorzystaniem endoskopu przewodowego lub bezprzewodowego. Pierwsza z technik jest zdecydowanie bardziej inwazyjna i zazwyczaj wymaga podania środków znieczulających. Drugą metodą jest endoskopia kapsułkowa. Polega ona na połknięciu przez pacjenta kapsułki o długości nie przekraczającej 31 mm i średnicy 11 mm, która jest wyposażona w miniaturową kamerę i układ do bezprzewodowej transmisji danych. Jest to badanie nieinwazyjne i umożliwia zbadanie wnętrza całego przewodu pokarmowego.

Bezprzewodowe kapsułki endoskopowe mają ograniczone zasoby energii dostępnej w miniaturowej baterii. Z tego powodu konieczne jest maksymalne zmniejszenie mocy nadajnika. W takim przypadku, parametry transmisji danych można poprawić wykorzystując zaawansowane techniki kodowania kanałowego [7] lub poprawiając parametry anteny. W drugim przypadku można zminimalizować straty odbiciowe anteny, które mogą być spowodowane efektem odstrojenia jej impedancji wejściowej w otoczeniu różnych tkanek ciała. Możliwe jest również zwiększenie zysku energetycznego anteny wpływającego na poziom mocy odebranej, a co za tym idzie na bilans energetyczny łącza pomiędzy nadajnikiem i odbiornikiem. Niestety, ze względu na bardzo ograniczone rozmiary anteny w kapsułce, nie jest możliwe znaczące zwiększenie jej zysku energetycznego. Kapsułka ma typowo kształt zbliżony do walca o wysokości od 24,5 do 31 mm i średnicy od 10,8 do 11,4 mm [5]. Antena mieszcząca się w tak małej objętości musi pracować z relatywnie długą falą odpowiadającą częstotliwości przeznaczonej dla implantowanych urządzeniach medycznych. W przypadku kapsułek endoskopowych jest to pasmo MICS (*ang. Medical Implant Communication Service*) o zakresie częstotliwości od 401 MHz do 406 MHz [10].

Pośród licznych konstrukcji anten endoskopowych można wyróżnić anteny wykorzystujące geometrię meandryczną i helikalną. Dzięki wykorzystaniu geometrii meandrycznej do realizacji promiennika anteny możliwe jest ograniczenie wymiarów anteny oraz wykonanie promiennika na elastycznym podłożu i ułożenie go w pobliżu ścianki kapsułki. Antena meandryczna opisana w [1] została zaprojektowana na fale krótsze niż w pasmie MICS, ponieważ pracuje w pasmie 2,40 GHz. Posiada strukturę przestrzenną (2 warstwy oddzielone dielektrykiem) i została wykonana na nieelastycznym podłożu Rogers 3010, co utrudnia dopasowanie tej konstrukcji do kapsułki endoskopowej. Inne anteny endoskopów wykorzystujące geometrię meandryczną [11, 6] mają budowę planarną i dzięki temu, że zostały wykonane na elastycznym laminacie można je ułożyć wewnątrz kapsułki. Ze względu na dużą wrażliwość impedancji wejściowej anteny opisanej w [11] na obiekty znajdujące się pobliżu szczeliny, w tym przypadku może dochodzić do odstrojenia anteny wynikającego z różnych właściwości elektrycznych poszczególnych tkanek otaczających antenę w danej części przewodu pokarmowego. Antena opisana w [6] została zaprojektowana w postaci dipola pętlowego, którego wymiary zostały ograniczone dzięki zastosowaniu sekcji meandrycznej. Antena ta jest jednak wrażliwa na otaczający materiał, ponieważ zmienia znacząco dopasowanie

impedancyjne dla różnych głębokości lokalizacji anteny w tkankach. W literaturze opisywane są również anteny endoskopów wykorzystujące geometrię spiralną [2], bądź helikalną [12]. W takim przypadku najczęściej są to anteny zajmujące pewną objętość kapsułki, z drugiej strony są to konstrukcje, które mogą być wykonane w łatwy i szybki sposób z cienkich przewodów.

W artykule przedstawiono projekt anteny przeznaczonej do pracy w endoskopie kapsułkowym w pasmie MICS 400 MHz wykorzystującej geometrię meandryczną. Opracowana antena charakteryzuje się dobrym dopasowaniem impedancyjnym w różnych obszarach przewodu pokarmowego człowieka. Konstrukcja anteny umożliwia dopasowanie jej kształtu do obudowy endoskopu kapsułkowego.

2. PROJEKT ANTENY

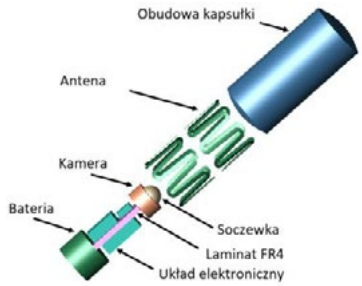
Założono, że projektowana antena będzie miała wymiary nieprzekraczające rozmiarów kapsułki endoskopowej (maksymalna wysokość anteny nie może przekroczyć 26 mm, jej średnica 11 mm) oraz będzie mogła być umieszczona na wewnętrznej powierzchni obudowy kapsułki.

Podczas projektowania anteny skupiono się na zapewnieniu jej jak najlepszego dopasowania impedancyjnego do impedancji 50 Ω w pasmie MICS 400 MHz. Podzespoły elektroniczne kapsułki takie jak kamera czy komponenty elektroniczne wpływają na dopasowanie impedancyjne anteny umieszczonej w ich pobliżu. W związku z tym na etapie projektowania anteny wykorzystano numeryczny model kapsułki przedstawiony na rys. 1. Składa się on z elementów o parametrach elektrycznych zbliżonych do parametrów materiałów wykorzystywanych do ich produkcji. W przypadku elementów elektronicznych przyjęto parametry glinki ceramicznej, soczewkę przybliżono parametrami szkła, baterię zamodelowano za pomocą idealnego przewodnika elektrycznego. W przypadku kamery przyjęto właściwości krzemu, a dla obwodu drukowanego przyjęto właściwości laminatu FR4 (ε_r = 4,1, $tg(\delta)$= 0,02). Cała kapsułka została zamknięta w obudowie z glinki ceramicznej o długości równej 26 mm i średnicy 11 mm, podobnie jak w przypadku innej konstrukcji konforemnej opisanej w literaturze [3]. Przestrzeń pomiędzy elementami wewnętrznymi kapsułki i obudową wypełniona jest żywicą epoksydową. Parametry zastosowanych materiałów zaczerpnięto z biblioteki materiałów programu XFdtd [9] i przedstawiono w tabeli 1.

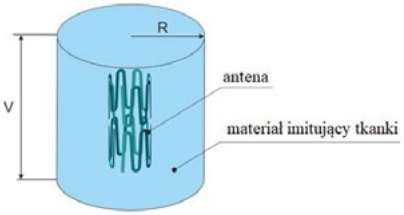
Symulacje komputerowe anteny przeprowadzono za pomocą programu XFdtd firmy Remcom wykorzystującego metodę różnic skończonych w dziedzinie czasu [9]. Wykorzystano uproszczony, jednorodny model ciała człowieka uśredniający parametry elektryczne tkanek (ε_r = 57, σ = 0,76 S/m). Parametry te są zgodne z właściwościami płynu symulującego tkanki ciała człowieka, który zostanie wykorzystany do eksperymentalnej weryfikacji anteny. Model miał kształt walca pokazanego na rys. 2: wysokość wynosi V =250 mm, a promień R =125 mm.

Tab. 1.Parametry elektryczne materiałów wykorzystanych do opracowania modelu numerycznego endoskopu

Materiał	Przenikalność elektryczna $\boldsymbol{\varepsilon_r}$	$tg(\delta)$
Glinka ceramiczna	9,8	0.0002
Szkło	3,8	0.0005
Żywica epoksydowa	4	0.04
Krzem	12,3	0.005

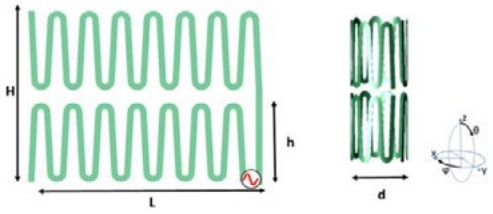


Rys. 1. Model kapsułki endoskopowej wykorzystany symulacjach



Rys. 2. Antena w modelu ciała imitującym tkanki

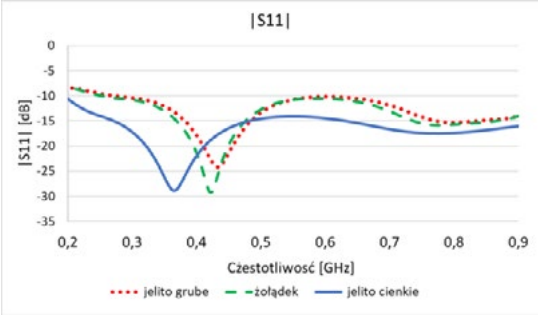
W wyniku przeprowadzonych symulacji opracowano konstrukcję anteny przedstawioną na rys. 3. Posiada ona promiennik meandryczny zbliżony kształtem do anteny typu „zig-zag" [4], który jest ułożony na powierzchni bocznej walca. Wymiary anteny wynoszą odpowiednio: H = 20 mm, a średnica d = 10 mm. W symulacjach zastosowano źródło napięciowe podłączone do dwu sekcji promiennika (rys. 3) oraz wykorzystano pobudzenie sygnałem Gaussowskim.



Rys.3. Promiennik opracowanej anteny endoskopowej: a – rozwinięcie na płaszczyźnie, b – widok po ułożeniu na powierzchni kapsułki.

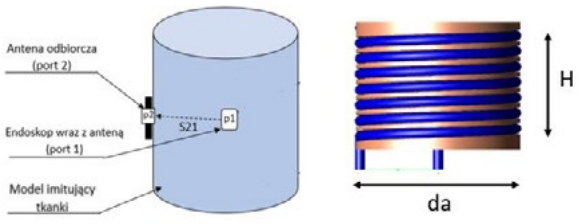
Kapsułka endoskopowa podczas badania przemieszcza się w przewodzie pokarmowym i jest otoczona tkankami, które różnią się miedzy sobą co ma wpływ na dopasowanie impedancyjne anteny. Z tego względu ważne jest żeby antena miała na tyle szerokie pasmo, aby pomimo wpływu różnych tkanek mogła pracować na częstotliwości 400 MHz. Zaprojektowaną antenę zweryfikowano umieszczając ją w modelu o właściwościach różnych tkanek, których parametry elektryczne również zaczerpnięto z biblioteki programu XFdtd [9]. Przyjęto, że pasmo impedancyjne anteny ograniczone jest przez

częstotliwości, dla których wartość $S_{11} \leq -10$ dB [1]. Wyniki przedstawione na rys. 4 pokazują, że zaproponowana antena cechuje się dobrym dopasowaniem impedancyjnym dla częstotliwości 400 MHz dla różnych tkanek wchodzących w skład przewodu pokarmowego ponieważ w założonym pasmie $S_{11} \leq -15$ dB.



Rys.4. Dopasowanie impedancyjne zaprojektowanej anteny dla różnych tkanek.

W zależności od położenia kapsułki organizm ludzki może tłumić falę emitowaną przez endoskop w różnym stopniu. W związku z tym porównano wartość mocy odbieranej w odbiorniku zewnętrznym zamodelowanym jako dipol skrócony dopasowany impedancyjne do 400 MHz dla położenia na powierzchni modelu ciała. Porównanie przeprowadzono dla kapsułkowej anteny helikalnej [12] przedstawionej na rys. 5b oraz zaprojektowanej anteny konforemnej „zig-zag". W przypadku anteny helikalnej zastosowano karkas wykonany w technologii druku 3D, który pozwolił na wykonanie fizycznego odpowiednika z zachowaniem wszystkich wymiarów. Symulacje przeprowadzono dla modelu uśrednionego oraz trzech modeli, których parametry elektryczne odpowiadają tkankom narządów układu pokarmowego (żołądek, jelito cienkie, jelito grube). Tłumienie sygnału między antenami symulowano wykorzystując parametry macierzy rozproszenia reprezentujące sprzężenie między portami anten zgodnie z układem pokazanym na rys. 5a.



a) b)

Rys. 5. a) układ do weryfikacji anten kapsułkowych wraz z definicja portów, b) referencyjna antena helikalna: H = 5,6 mm, da = 8 mm [12].

W tabeli 2 przedstawiono średnie wartości parametrów S_{21} uzyskane dla 5 różnych pozycji anteny kapsułkowej. Podczas symulacji odległość między antenami umieszczonymi na tej samej wysokości zmieniano w zakresie od 10 mm do 240 mm co 64 mm. Zaproponowana antena konforemna zapewnia mniejsze tłumienie sygnału pomiędzy anteną wewnętrzna i antenami zewnętrznymi w porównaniu z anteną helikalną. Różnica waha się od około 3 dB do 7 dB dla różnych położeń anteny kapsułkowej.

Tab. 2. Średnie tłumienie sygnału między antenami (parametry S) dla różnych tkanek

	Antena helikalna $\overline{S_{21}}$ [dB]	**Antena konforemna „zig-zag" $\overline{S_{21}}$ [dB]**
Model uśredniony	-68	-65
Jelito grube	-73,5	-70
Jelito cienkie	-84,4	-80,7
Żołądek	-72,5	-69,5

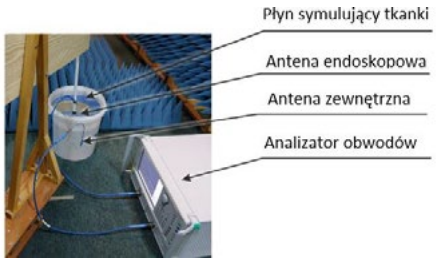
3. WERYFIKACJA POMIAROWA

Obudowę anteny prototypowej imitującą kapsułkę wykonano za pomocą technologii druku 3D. W tym celu wykorzystano metodę osadzania topionego materiału (*ang. Fused Deposition Modeling (FDM)*) z wykorzystaniem materiału, którym był polimer PLA o przenikalności elektrycznej wynoszącej ε_r=4,1 i stratności dielektrycznej $tg(\sigma)$=0,01. W przypadku, zastosowania materiału PLA skompensowano jego wypływ przez nieznacznie zwiększenie wysokość promiennika anteny o 3 mm (H= 23 mm). Promiennik anteny jest wykonany z laminatu polimidowego Pyralux firmy DuPont, o grubości 25 µm pokrytego miedzią o grubości 45 µm. Przenikalność elektryczna podłoża podana przez producenta to ε_r = 3,4 a stratność dielektryczna $tg(\delta)$ = 0,005. Ponieważ materiał jest warstwą znacznie cieńszą od wielkości woksela (w tym przypadku 0,5 mm) to przyjęto, że jego efektywna przenikalność wykorzystana do symulacji numerycznej wynosi ε_r = 1,7, a strata dielektryczna $tg(\delta)$ = 0,001 [8]. Promiennik anteny przedstawiono na rysunku 6. Do jego wykonania zastosowano proces trawienia chemicznego. Promiennik został przylutowany bezpośrednio do kabla koncentrycznego umożliwiającego przeprowadzenie pomiarów (nie zastosowano symetryzatora). Aby ułożyć promiennik na powierzchni bocznej kapsułki i ustabilizować jego kształt, został on zwinięty i umieszczony pomiędzy dwoma elementami cylindrycznymi wydrukowanymi w z materiału PLA, jak to pokazano na rys. 6b.

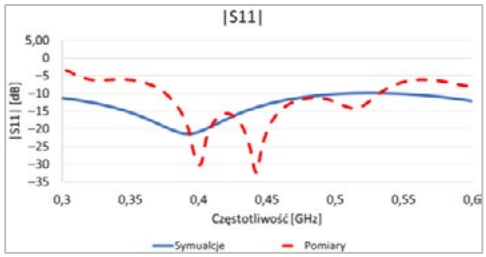
Opracowaną antenę zweryfikowano za pomocą pomiarów przeprowadzonych z wykorzystaniem wektorowego analizatora obwodów oraz uproszczonego modelu ciała. Model ten przedstawiony na rys. 7, wykonano z cienkościennego naczynia z tworzywa sztucznego wypełnionego płynem imitującym tkanki. Wyniki pomiarów dopasowania impedancyjnego anteny pokazane na rys. 8 pokazują, że opracowana antena może być wykorzystana w pasmie 400. Wyniki pomiarów tłumienia między anteną kapsułkową i anteną zewnętrzną zaprezentowano na rys.9. W przypadku wykorzystania zaprojektowanej anteny tłumienie sygnału jest o około 9 dB mniejsze w porównaniu do anteny helikalnej.

a b

Rys. 6. Prototyp opracowanej anteny meandrycznej „zig-zag": (a) promiennik anteny wykonany na podłożu Pyralux, (b) promiennik anteny podczas wsuwania go w osłonę.



Rys. 7. Model uproszczony ciała człowieka wykorzystany do eksperymentalnej weryfikacji anteny kapsułkowej



Rys. 8. Dopasowanie impedancyjne anteny uzyskane w symulacjach i pomiarach dla modelu uśrednionego.

Rys. 9. Wyniki pomiaru parametru S21 dla anteny helikalnej oraz prototypowej anteny "zig-zag"

4. PODSUMOWANIE

Opracowana antena endoskopu typu ‘zig-zag’ charakteryzuje się zadowalającym dopasowaniem impedancyjnym w pasmie MICS 400 MHz. Jak pokazano na rys. 4, w tym paśmie parametr S_{11} nie przekracza wartości -15 dB w przypadku usytuowania anteny we wszystkich rozważanych tkankach. W przypadku zastosowania tej anteny, średnie tłumienie sygnału między anteną kapsułki i anteną odbiorczą jest mniejsze niż w przypadku opisanej w literaturze anteny helikalnej. Dla przyjętego środowiska symulacyjnego uzyskano poprawę do około 7 dB. Zastosowanie uproszczonego modelu ciała człowieka jest dużym uproszczeniem, jednak w przypadku nadajników, które zmieniają swoje położenie we wnętrzu ciała przyjęcie takiego modelu wydaje się być uzasadnione ze względu na duża zmienność parametrów poszczególnych tkanek. Ponadto, taki model można odtworzyć w warunkach eksperymentu pomiarowego.

Ze względu na swoją konstrukcję opracowana antena nadaje się do zastosowania w kapsułkach endoskopowych ponieważ może być wykonana z wykorzystaniem elastycznego laminatu cienkowarstwowego. Ponadto można ją umieścić na powierzchni obudowy kapsułki co pozwala zachować przestrzeń wewnętrzną dla innych elementów elektronicznych.

LITERATURA

[1] Feng, Yang, Zhaonan Li, Lin Qi, Wanting Shen, and Gaosheng Li. 2022. “*A Compact and Miniaturized Implantable Antenna for ISM Band in Wireless Cardiac Pacemaker System*”. Scientific Reports 12 (1): 238.

[2] Lee, Sang Heun, Kihun Chang, Ki Joon Kim, and Young Joong Yoon. 2008. “*A Conical Spiral Antenna for Wideband Capsule Endoscope System*”. In 2008 IEEE Antennas and Propagation Society International Symposium, 1–4.

[3] Abdul Basir, Muhammad Zada, Youngdae Cho, and Hyoungsuk Yoo. 2020. “*A Dual-Circular-Polarized Endoscopic Antenna With Wideband Characteristics and Wireless Biotelemetric Link Characterization*”. IEEE Transactions on Antennas and Propagation 68 (10): 6953–63.

[4] Sing Lee, and K. Mei. 1970. “*Analysis of Zigzag Antennas*”. IEEE Transactions on Antennas and Propagation 18 (6): 760–64.

[5] Ionescu, Alin, Adina Glodeanu, Mihaela Ionescu, Sorin Zaharie, Ana Ciurea, Andreea Golli, Nikolaos Mavritsakis, Didi Popa, and Cristin Vere. 2022. “*Clinical Impact of Wireless Capsule Endoscopy for Small Bowel Investigation (Review)*”. Experimental and Therapeutic Medicine 23 (4): 262.

[6] Alemaryeen, Ala. 2021. “*Compact Wideband Antenna for Wireless Capsule Endoscopy System*”. Applied Physics A 127 (4): 271.

[7] Mezher, Mohanad Ahmed, Sadia Din, Muhammad Ilyas, Oguz Bayat, Qammer Hussain Abbasi, and Imran Ashraf. 2022. *“Data Transmission Enhancement Using Optimal Coding Technique Over In Vivo Channel for Interbody Communication*”. *Big Data*, April, big.2021.0224.

[8] Januszkiewicz, Łukasz, Paolo Di Barba, and Jarosław Kawecki. 2021. “*Design Optimization of Wearable Multiband Antenna Using Evolutionary Algorithm Tuned with Dipole Benchmark Problem*”. Electronics 10 (18): 2249.

[9] Reference Manual-XFdtd. Dostęp on-line: https://support.remcom.com/xfdtd/reference.html

[10] ITU-R RS. 1346. 1998. “*Sharing Between the Meteorological Aids Service and Medical Implant Communication Systems (MICS) Operating in the Mobile Service in the Frequency Band 401-406 MHz*”.

[11] Bao, Zengdi, Yong-Xin Guo, and Raj Mittra. 2017. “*Single-Layer Dual-/Tri-Band Inverted-F Antennas for Conformal Capsule Type of Applications*”. IEEE Transactions on Antennas and Propagation 65 (12): 7257–65.

[12] Kwak, Sang Il, Kihun Chang, and Young Joong Yoon. 2005. “*The Helical Antenna for the Capsule Endoscope*”. In 2005 IEEE Antennas and Propagation Society International Symposium, 2B:804–7 vol. 2B.

KONCEPCJA ZDALNEGO STEROWANIA BLOKU WZMACNIACZY ANTENOWYCH SYSTEMU ANTENOWEGO AM 524 W KOMORACH SAC

Leszek Nowosielski[1]; Bartosz Dudziński[2]; Michał Nowosielski[3]; Aleksandra Maria Ślubowska[4]
[1] Wojskowa Akademia Techniczna, Wydział Elektroniki ul. Kaliskiego 2, Warszawa, leszek.nowosielksi@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Wydział Elektronikiul. Kaliskiego 2, Warszawa, bartosz.dudzinski@wat.edu.pl
[3] Uniwersyteckie Centrum Kliniczne Warszawskiego Uniwersytetu Medycznego, Warszawa, M.Nowosielski@outlook.com
[4] Uniwersyteckie Centrum Kliniczne Warszawskiego Uniwersytetu Medycznego, Warszawa, A.Slubowska@outlook.com

DOI: 10.15199/59.2023.4.69

Streszczenie: **Do pomiarów emisji ujawniającej firma Rohde & Schwarz wyprodukowała system czułych anten aktywnych AM524 zintegrowany z zestawem wzmacniaczy antenowych: GX 525, GX 526 oraz GX 527. Powyższy system antenowy jest powszechnie stosowany we współczesnych laboratoriach badawczych zajmujących badaniami ulotu informacji z urządzeń informatycznych. W artykule opisano koncepcję układu zdalnego sterowania pracą bloku wzmacniaczy antenowych systemu antenowego AM 524 wykorzystywanego podczas badań ulotu informacji z urządzeń informatycznych. System ten ma za zadanie umożliwić zdalne przełączanie pięcioma wejściami znajdującymi się na wzmacniaczach antenowych: GX 525, GX 526 oraz GX 527 z poziomu komputera PC wykorzystującego odpowiednie oprogramowanie. Autorzy artykułu opisali system antenowy AM 524. W artykule zawarto również opis koncepcji układu zdalnego sterowania, a także modułu ADAM 6052 stanowiącego centralny element urządzenia.**
Abstract: **For the analysis of the emission of publications, Rohde & Schwarz produced a system of sensitive antennas operating AM524, partly with a set of antenna amplifiers: GX 525, GX 526 and GX 527. The transmitted antenna system is modern in modern research laboratories along with research on information leakage from devices. In the document created as part of the remote control system for the operation of the antenna amplifier block of the AM 524 antenna system, information leakage from device devices is used during testing. This system is designed to connect with five inputs on the antenna amplifiers: GX 525, GX 526 and GX 527 from the level of the PC included with the software. The authors of the article described the AM 524 antenna system. The article also describes the functions of the remote control system, as well as the ADAM 6052 module of a specific element of the central device.**

Słowa kluczowe: **EMC, infiltracja elektromagnetyczna, zdalne sterowanie**

Keywords: **EMC, electromagnetic infiltration, remote control**

1. WSTĘP

Niewiarygodnie szybki rozwój techniki, jaki obserwujemy w ostatnich latach, a w szczególności rozwój technik telekomunikacyjnych i lawinowa komputeryzacja niemal każdej dziedziny życia sprawiają, że obecnie najcenniejszym towarem jest informacja. To właśnie ona jest kluczem do sukcesu zarówno w polityce, jak i w biznesie. Ponieważ znakomita większość informacji przetwarzana jest obecnie z wykorzystaniem różnego rodzaju urządzeń informatycznych, dużego znaczenia nabiera problematyka ich zabezpieczenia przed nieuprawnionym przejęciem. Duże zagrożenie w tym zakresie stwarzają emisje zaburzeń elektromagnetycznych generowane nieintencjonalnie przez urządzenia informatyczne. W sprzyjających warunkach możliwe jest wykorzystanie tej niepożądanej emisji zaburzeń do odtworzenia przez nieupoważnione osoby przetwarzanej w urządzeniu informacji. Powyższy proces określany jest jako infiltracja elektromagnetyczna. Emisję elektromagnetyczną, której sygnał skorelowany jest z informacją użyteczną i może być wykorzystany do jej odtworzenia, nazywamy emisją ujawniającą. Emisja ujawniająca umożliwiająca infiltrację elektromagnetyczną jest nazywana elektromagnetycznym przenikaniem informacji.

Do badania emisji ujawniającej firma Rohde & Schwarz wyprodukowała system bardzo czułych anten aktywnych AM524 zintegrowany z zestawem wzmacniaczy antenowych: GX 525, GX 526 oraz GX 527. Powyższy system antenowy jest powszechnie stosowany we współczesnych laboratoriach badawczych zajmujących badaniami ulotu informacji z urządzeń informatycznych. Standardowo system anten aktywnych AM524 jest sprzedawany przez producenta razem z systemem zdalnego sterowania bazującym na medium transmisyjnym komend zdalnego sterowania jakim jest światłowód. Niestety technika przesyłania komend zdalnego sterowania za pośrednictwem światłowodów posiada szereg wad:

- wymaga instalowania specjalnych światłowodowych adapterów panelowych w panelu wprowadzeń liniowych komory bezodbiciowej SAC, które mogą prowadzić do znacznej redukcji wartości skuteczności ekranowania całej komory,
- światłowody cechują się dużą wrażliwością na uszkodzenia mechaniczne,
- naprawa łączy światłowodowych jest dosyć kosztowna i wymaga specjalizowane aparatury serwisowej.

W niniejszym artykule przedstawiono opis koncepcji budowy systemu zdalnego sterowania zestawem wzmacniaczy antenowych: GX 525, GX 526 oraz GX 52,

które bazuje zna medium transmisyjnym jakim są przewody koncentryczne i nie posiada wyżej wymienionych wad systemu bazującego na światłowodach.

Wyposażenie pomiarowe we współczesnych laboratoriach zajmujących się badaniami ulotu informacji z urządzeń informatycznych jest instalowane w dwóch sąsiednich pomieszczeniach, odseparowanych od siebie szczelną barierą elektromagnetyczną. Jednym z pomieszczeń jest komora bezodbiciowa SAC, której podstawową funkcją jest zapewnienie izolacji badanego urządzenia od zewnętrznego środowiska elektromagnetycznego. Izolację powyższą zapewnia bariera elektromagnetyczna, będąca integralną częścią komory bezodbiciowej, w skład której wchodzi obudowa ekranująca cechująca się dużą wartością skuteczności ekranowania (około 100 dB) oraz filtry sygnałowe i zasilania. Izolacja badanego urządzenia od zewnętrznego środowiska elektromagnetycznego, w trakcie przeprowadzania pomiarów emisyjności urządzeń, umożliwia odseparowanie emisji zaburzeń generowanych przez obiekt badań, które są przedmiotem pomiarów, od zewnętrznych zaburzeń środowiskowych. W tym celu w komorach bezodbiciowych umieszcza się jedynie obiekt badań oraz sensory pomiarowe takie jak: anteny, sieci sztuczne, cęgi absorpcyjne. Pozostałe wyposażenie (np. odbiornik pomiarowy, komutator obwodów w.cz.) niezbędne do przeprowadzenia pomiarów poziomu emisji zaburzeń promieniowanych lub przewodzonych, które mogą być potencjalnym źródłem dodatkowych emisji zaburzeń, umieszcza się w sąsiednim pomieszczeniu przyrządowym.

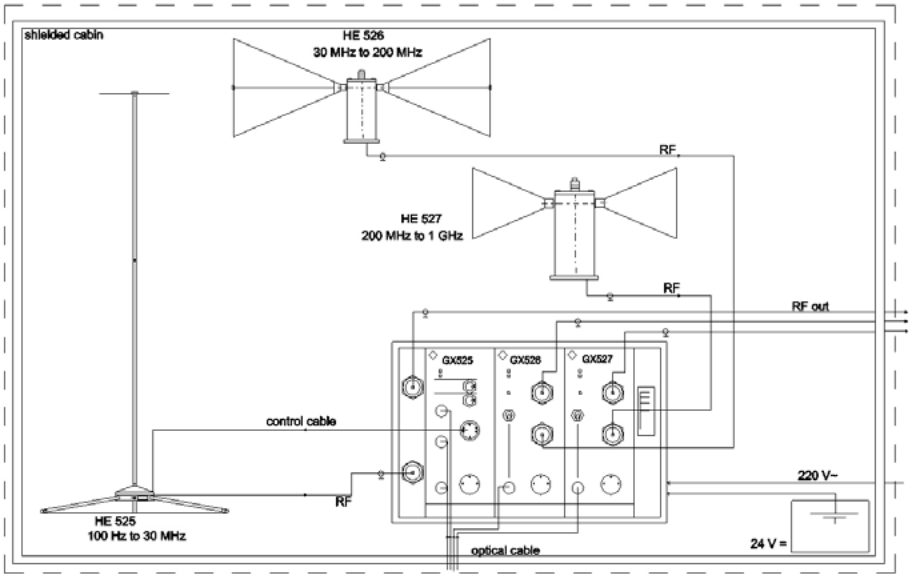
W celu skrócenia badań oraz właściwego skonfigurowania układu do pomiaru poziomu emisji niezbędne jest połączenie wyjścia odpowiedniego sensora pomiarowego do wejścia miernika poziomu sygnału wysokiej częstotliwości. W przypadku pomiaru poziomu emisji zaburzeń promieniowanych do wejścia miernika poziomu sygnału podłącza się anteny pomiarowe pracujące w różnych zakresach częstotliwości. W praktyce przełączanie tzw. szynku antenowego np. aktywnego systemu antenowego AM524 odbywa się przy pomocy jednostki kontrolującej GS525, sterowanej za pomocą przewodów optycznych GS525K1 (rys. 1).

Niskoszumowy aktywny system antenowy AM524 został zaprojektowany do pomiaru sygnałów o niskim poziomie w komorach bezodbiciowych SAC. Zakres częstotliwości pracy anten wynosi od 100 Hz do 1000 MHz i jest podzielony na trzy podzakresy:

od 100 Hz do 30 MHz antena prętowa HE 525,
od 30 MHz do 200 MHz antena dipolowa HE 526,
od 200 MHz do 1000 MHz antena dipolowa HE 527.

Złącza modułów: GX 525, GX 526 oraz GX 527 (rys.1) wchodzące w skład systemu antenowego AM 524 zapewniają zasilanie dla anten wchodzących w skład systemu oraz umożliwiają sterowanie systemem antenowym.

W niniejszym artykule autorzy zaproponowali koncepcję innowacyjnej metody zdalnego sterowania blokami wzmacniaczy antenowych: GX 525, GX 526 oraz GX 527 (rys.1) za pośrednictwem przewodów koncentrycznych zamiast proponowanych przez producenta światłowodów. W zaproponowanym rozwiązaniu zdalnego sterowania sygnały sterujące są przesyłane za pośrednictwem przewodu koncentrycznego (przewód gorący oraz ekran przewodu koncentrycznego) w formie napięcia zasilającego diodę LED umieszczoną na końcu przewodu. Sterowanie poszczególnymi wejściami wzmacniaczy antenowych odbywa się przy pomocy impulsów świetlnych generowanych przez diodę LED zainstalowaną na końcu każdego z przewodów koncentrycznych. Przewaga opracowanego rozwiązania zdalnego sterowania blokami wzmacniaczy antenowych, w stosunku do fabrycznego rozwiązania zdalnego sterowania wykorzystującego światłowody, polega na braku konieczności instalowania na panelu wejść liniowych komory SAC dodatkowych światłowodowych adapterów panelowych. Zamiast powyższych adapterów można wykorzystać standardowo instalowane w komorach SAC koncentryczne adaptery panelowe służące do przesyłania sygnałów w.cz.

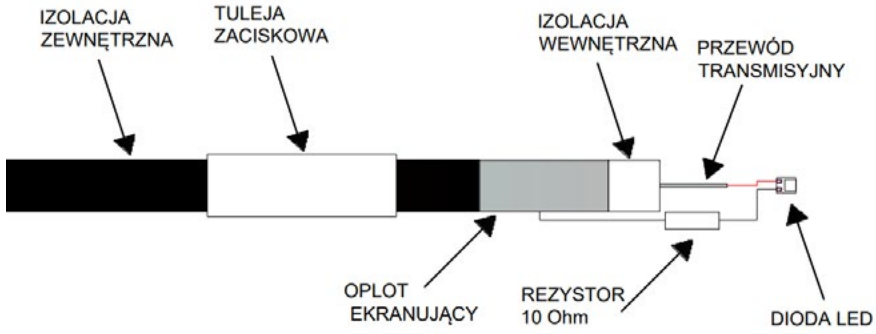


Rys. 1. Konfiguracja aktywnego systemu antenowego AM 524

2. KONCEPCJA SYSTEMU ZDALNEGO STEROWANIA PRACĄ BLOKU WZMACNIACZY ANTENOWYCH AM 524

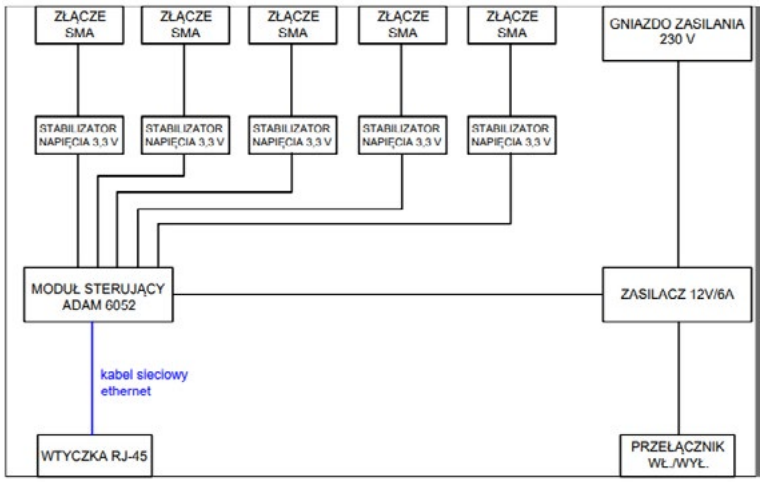
Poniższy rozdział zawiera opis koncepcji budowy systemu zdalnego sterowania pracą wzmacniaczy antenowych: GX 525, GX 526 oraz GX 527 z poziomu komputera PC. Ogólny schemat systemu zdalnego sterowania wzmacniaczami antenowymi przedstawiono na rys. 2. W skład systemu zdalnego sterowania wzmacniaczami antenowymi wchodzi: pięć zmodyfikowanych przewodów koncentrycznych, układ zdalnego sterowania oraz oprogramowanie zdalnego sterowania zainstalowane na komputerze typu PC.

Rys. 2. Ogólny schemat systemu zdalnego sterowania wzmacniaczami antenowymi



Rys. 3. Zmodyfikowany przewód koncentryczny

Budowę zmodyfikowanego przewodu koncentrycznego zdalnego sterowania przedstawiono na rys. 3. Modyfikacja przewodu koncentrycznego polega na zainstalowaniu na jednym z końców przewodu diody LED razem z rezystorem ograniczającym przepływ prądu przez diodę. Katodę diody LED należy podłączyć do przewodu centralnego kabla koncentrycznego, natomiast anoda diody LED wraz z rezystorem węglowym o rezystancji 10 Ω należy podłączyć do oplotu ekranującego kabla. Zastosowanie rezystora węglowego ograniczy wartości prądu zasilającego diodę LED do 58 mA, dzięki czemu nagrzewanie się połączenia lutowniczego zostanie ograniczone. Dioda LED wraz z rezystorem węglowym zostanie umieszczona wewnątrz złącza SMA typu męskiego. Zadaniem przewodu koncentrycznego jest dostarczenie z układu zdalnego sterowania napięcia zasilającego diodę LED o wartości 3,3 V. Potencjał dodatni zasilania diody LED zostanie przesłany przewodem gorącym kabla koncentrycznego natomiast masa będzie podłączona do oplotu ekranującego przewód koncentryczny.



Rys. 4. Schemat blokowy układu zdalnego sterowania

Schemat blokowy układu zdalnego sterowania przedstawiono na rys. 4. Najważniejszym komponentem układu zdalnego sterowania jest moduł sterujący typu ADAM 6052. Widok zewnętrzny modułu sterującego ADAM 6052 przedstawiono na rys. 5. Moduł powyższy umożliwia integrację systemów automatyki poprzez wykorzystanie technologii Internetu. Moduł ten stanowi centralny element układu zdalnego sterowania, który umożliwia zamianę komend zdalnego sterowania wzmacniaczy antenowych, przesyłanych przy pomocy protokołu TCP/IP z komputera PC, generowanych przez oprogramowanie zdalnego sterowania, na sygnały stałonapięciowe 0 V / 12 V podawane na wyjścia cyfrowe modułu sterującego ADAM 6052. W zaprojektowanym układzie wykorzystano pięć wyjść modułu ADAM 6052, z których sygnały stałonapięciowe 0 V / 12 V są przekazywane do układów stabilizacji napięcia o wartości 3,3 V koniecznej do zapewnienia odpowiednich warunków pracy diody LED umieszczonej w zmodyfikowanym przewodzie koncentrycznym. Poszczególne zmodyfikowane przewody koncentryczne są podłączone z jednej strony do jednego z wejść sterowanych wzmacniaczy antenowych, a z drugiej strony do jednego ze złącz SMA układu zdalnego sterowania. Wszystkie elementy urządzenia zostaną umieszczone w obudowie wykonanej z tworzywa sztucznego.

Komputer PC należy podłączyć do układu zdalnego sterowania przy pomocy przewodu Ethernet.

Dioda LED stanowić będzie źródło światła barwy białej. Pod wpływem takiego impulsu świetlnego, nastąpi reakcja układu wzmacniacza antenowego (jego przełączenie ze stanu wyłączenia na stan włączenia).

Wykorzystanie kabla koncentrycznego oraz diody LED umożliwia zastąpienie niskim kosztem drogiego, dedykowanego przewodu światłowodowego. Dodatkowo, przewód współosiowy spełnia wymagania szczelności elektromagnetycznej oraz cechuje się wysoką odpornością na uszkodzenia fizyczne.

Rys. 5. Widok zewnętrzny modułu sterującego ADAM 6052

3. WNIOSKI

W artykule autorzy zaproponowali koncepcję innowacyjnej metody zdalnego sterowania blokami wzmacniaczy antenowych: GX 525, GX 526 oraz GX 527 za pośrednictwem przewodów koncentrycznych zamiast proponowanych przez producenta światłowodów. W zaproponowanym rozwiązaniu zdalnego sterowania sygnały sterujące są przesyłane za pośrednictwem przewodu koncentrycznego (przewód gorący oraz ekran przewodu koncentrycznego) w formie napięcia zasilającego diodę LED umieszczoną na końcu przewodu. Sterowanie poszczególnymi wejściami wzmacniaczy antenowych odbywa się przy pomocy impulsów świetlnych generowanych przez diodę LED zainstalowaną na końcu każdego z przewodów koncentrycznych.

Przewaga opracowanego rozwiązania zdalnego sterowania blokami wzmacniaczy antenowych, w stosunku do fabrycznego rozwiązania zdalnego sterowania wykorzystującego światłowody, polega na braku konieczności instalowania na panelu wejść liniowych komory SAC dodatkowych światłowodowych adapterów panelowych. Zamiast powyższych adapterów można wykorzystać standardowo instalowane w komorach SAC koncentryczne adaptery panelowe służące do przesyłania sygnałów w.cz.

Opracowana koncepcja systemu zdalnego sterowania pracą bloku wzmacniaczy antenowych systemu antenowego AM 524 umożliwi zdalne sterowanie wzmacniaczy antenowych: GX 525, GX 526 oraz GX 527 z poziomu komputera PC wykorzystującego dedykowane do tego celu oprogramowanie.

LITERATURA

[1] http://led.gliwice.pl/-ogolnego-stosowania-ip20/52-zasilacz-led-60-w.html
[2] https://www.kolbinger.at/DAQM/manuals/UM-ADAM-6000-Ed7-1EN_2015_09.pdf
[3] https://pdf1.alldatasheet.com/datasheet-pdf/view/49118/ADMOS/AMS1117.html
[4] https://pl.aliexpress.com/item/32520903460.html
[5] ROHDE & SCHWARZ Documentation „Low noise active antenna system AM 524."
[6] https://www.rohde-schwarz.com/us/product/am524-productstartpage_63493-7871.html
[7] https://dam-assets.fluke.com/s3fs-public/151718__umeng0000.pdf
[8] http://krysiak.info/sieci/3ethernet.htm#3.7
[9] https://www.conrad.pl/p/modul-io-advantech-adam-6052-be-modbus-tcp-ea-16-kanalowy-197656
[10] https://www.advantech.com/products/a67f7853-013a-4b50-9b20-01798c56b090/adam-6052/mod_8e8d8d87-f30b-4f82-b77e-4e3822dfca5d
[11] https://www.elmark.com.pl/sklep/advantech/adam-6052-modul-pomiarowy-8di-8do-modbus-tcp
[12] https://www.dipol.com.pl/kompatybilnosc_elektromagnetyczna_bib53.htm
[13] https://automatykab2b.pl/temat-miesiaca/43297-anatomia-kabla-przemyslowego/strona/1-ekranowanie-czyli-o-warstwe-glebiej

Funding: This work was financed by Military University of Technology under research project UGB-22-859/2023 on "Estimation of electromagnetic compatibility between DVBT2 and 4G/5G in the 700 MHz band".

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

IMPLEMENTACJA METODY UKRYWANIA DANYCH W OBRAZACH Z WYKORZYSTANIEM SIECI GENERATYWNYCH

IMPLEMENTATION OF A METHOD FOR HIDING DATA IN IMAGES USING GENERATIVE NETWORKS

Tomasz Walczyna[1]; Zbigniew Piotrowski[2];

[1] Wojskowa Akademia Techniczna, Warszawa, tomasz.walczyna@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Warszawa, zbigniew.piotrowski@wat.edu.pl

DOI: 10.15199/59.2023.4.70

Streszczenie: **W artykule przedstawiamy metodę ukrywania danych w obrazach z użyciem sieci generatywnych - GAN (Generative Adversarial Networks), do generowania obrazów zawierających ukryte dane. Proponowana metoda pozwala na efektywne ukrywanie danych w obrazach, co może znaleźć zastosowanie w wielu dziedzinach, takich jak ochrona prywatności, steganografia, czy znakowanie wodne. Przedstawione w artykule wyniki stanowią podstawę do dalszych badań nad ulepszaniem metody i jej zastosowaniem w praktyce.**

Abstract: **This paper presents a method for hiding data in images using Generative Adversarial Networks (GAN) to generate images with secret data. The proposed method makes it possible to hide data in images effectively, which can find applications in many fields, such as privacy protection, steganography, and watermarking. The results presented in the article provide a basis for further research on improving the method and its application in practice.**

Słowa kluczowe: **sieci generatywne, watermarking, steganografia, sieci neuronowe**

Keywords: **generative networks, watermarking, steganography, neural networks**

1. WSTĘP

W dzisiejszych czasach, kiedy obrazy są tak powszechne i łatwo dostępne, zwiększa się potrzeba zapewnienia ich bezpieczeństwa i prywatności. Jednym z rozwiązań jest ukrywanie danych w obrazach, czyli watermarking. Jednakże, większość istniejących klasycznych metod watermarkingowych może być łatwo naruszona poprzez zakłócenia, takie jak m.in. szum, zmiana parametrów (jasność, kontrast, saturacja), blur, kompresję, czy wycinanie [6].

W tym artykule opisujemy metodę ukrywania ciągu bitów w obrazie, która jest odporna na powyższe zakłócenia. W tym celu wykorzystujemy połączenie dwóch potężnych narzędzi w dziedzinie sztucznej inteligencji: sieci U-net [7] oraz Generative Adversarial Network [1] (GAN).

Sieć U-net jest powszechnie stosowaną siecią konwolucyjną w dziedzinie przetwarzania obrazów, która została zaprojektowana do segmentacji obrazów medycznych. Natomiast GAN to model generatywny, który składa się z dwóch sieci neuronowych: generatora i dyskryminatora. GAN jest w stanie nauczyć się generowania nowych obrazów, które są niemal identyczne do prawdziwych obrazów, co sprawia, że może być wykorzystany w procesie ukrywania danych.

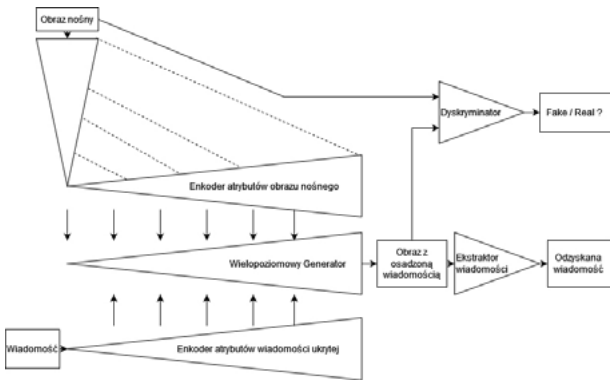
W naszej metodzie, wykorzystujemy sieć U-net jako ekstraktor cech nośnika oraz generator wiadomości o różnych rozmairach. Następnie cechy te są odpowiednio łączone i generowany jest obraz zawierający ukryte dane. Następnie, ten obraz jest wprowadzany do dyskryminatora GAN, który jest w stanie rozpoznać, czy obraz został zmieniony w taki sposób, aby ukryć dane. Ostatecznie, ciąg bitów jest pozyskiwany z obrazu nośnego przy pomocy ekstraktora opartego o enkoder.

Nasza metoda została przetestowana na różnych zakłóceniach, takich jak kompresja jpg, szum gaussowski, zasłanianie fragmentu obrazu, zmiana paramterów. Wyniki pokazują, że nasza metoda jest odporna na te zakłócenia i skutecznie ukrywa ciąg bitów w obrazie.

W tym artykule opisujemy naszą metodę i prezentujemy wyniki testów dla różnych pojemności, co może stanowić cenne wskazówki dla osób zajmujących się bezpieczeństwem i prywatnością obrazów.

2. METODA

Proponowana metoda składa się z pięciu pomniejszych modeli: enkoder atrybutów obrazu nośnego, enkoder atrybutów wiadomości ukrytej, wielopoziomowy generator, ekstraktor wiadomości oraz dyskryminator. Każdy z tych modeli jest odpowiedzialny za konkretne zadanie, które łącznie tworzą efektywny system ukrywania danych w obrazach.



Rys. 1. Struktura sieci

Zadania poszczególnych modeli:

- enkoder atrybutów obrazu nośnego – U-Net odpowiedzialny za ekstrakcję wielopoziomowych atrybutów nośnika;
- enkoder atrybutów wiadomości ukrytej – odpowiedzialny za adaptację wiadomości do wymiarów generowanego obrazu;
- wielopoziomowy generator – odpowiedzialny za konkatenację pozyskanych atrybutów i generację obrazu
- ekstraktor wiadomości – dekoder odpowiedzialny za pozyskanie wiadomości ukrytej z nośnika
- dyskryminator – model stosowany w trakcie treningu, odpowiedzialny za ocenę stopnia realizmu wygenerowanego obrazu.

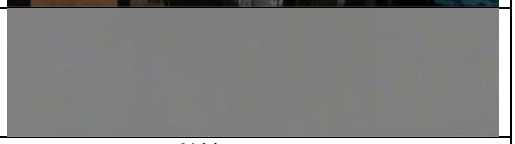
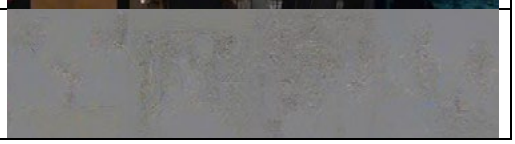
W ramach eksperymentów przyjęty został stały wymiar obrazu – 256x256. Enkoder atrybutów obrazu nośnego jest modelem sieci U-net, który pozyskuje wielowymiarowe atrybuty od wymiaru 4x4 do 256x256. Najmniejsza przestrzeń ukryta jest również podawana jako wejście dla generatora. W ramach enkodera atrybutów wiadomości ukrytej zastosowano zwykłe nadpróbkowanie/podpróbkowanie (w zależności od długości wiadomości) połączone z sieciami konwolucyjnymi, tak aby dopasować wymiary wiadomości do wymiarów obrazów. W ramach implementacji generatora sugerowano się dwoma modelami: FaceShifter [2] – wprowadzający bloki umożliwiające połączenia atrybutów z różnymi wagami określonymi przez maskę sigmoidalną oraz FastGAN [3] wykorzystany również w dyksryminatorze zapewniający stabilny trening sieci generatywnych. W ramach ekstraktora wiadomości ukrytej zastosowano prosty enkoder CNN-FCN.

W metodzie tej wykorzystano GAN - sieć neuronową składającą się z dwóch modeli: generatora i dyskryminatora. Generator służy do generowania fałszywych obrazów, podczas gdy dyskryminator ma za zadanie rozróżnianie prawdziwych obrazów od tych fałszywych. Obie te sieci konkurują ze sobą w trakcie treningu, aż do momentu, gdy generator zaczyna generować obrazy, które dyskryminator nie jest w stanie odróżnić od prawdziwych.

3. WYNIKI

W trakcie eksperymentów sprawdzono skuteczność algorytmu dla kluczy o różnych długościach: 16, 64, 256 i 1024 bitów. Wyniki testów były różne w zależności od długości klucza oraz rodzaju szumu, na którym był przeprowadzany test. Przykładowe próbki dla różnych scen ze zbioru testowego przedstawiono w tabeli 1. W wierszu „Watermark" przedstawiona jest różnica pomiędzy zdjęciem bez znaku (Cover), a tym już z ukrytym znakiem wodnym (Hidden). Podczas, gdy dla krótkiej wiadomości różnica ta jest ledwie zauważalna, dla dłuższych rośnie ona znacząco. Percepcyjnie nie ma to jednak wpływu na odbiór sceny.

Tab. 1. Wyniki ukrywania dla różnych pojemności znaku wodnego

Cover	
16 bitów	
Hidden	
Watermark	
64 bity	
Hidden	
Watermark	
256 bitów	
Hidden	
Watermark	
1024 bity	
Hidden	
Watermark	

Testy przeprowadzono na 80000 próbkach dla różnych rodzajów szumów. Jednakże, bardziej interesujące są wyniki uzyskane w tabeli 2, która przedstawia średni PSNR oraz SSIM dla każdej z 4 metod. PSNR (Peak Signal to Noise Ratio) oraz SSIM (Structural Similarity Index) to miary jakości ukrytego znaku wodnego. Można zauważyć, że im dłuższy klucz, tym mniejsza jakość ukrytego znaku wodnego. W przypadku mniejszej pojemności, bez manipulacji wagami, sieć sama wytrenowała się w taki sposób, by znak wodny był mniej widoczny.

Tab. 2. PSNR oraz SSIM dla różnych pojemności znaku wodnego

Bity	*PSNR*	*SSIM*
16	36,92	0,986
64	35,5	0,975
256	33,69	0,967
1024	31.96	0.957

W tabeli 3 przedstawiono średni BER przy zastosowaniu różnych rodzajów szumów. Jak można zauważyć najbardziej destrukcyjnym szumem była kompresja JPEG, zmiana jasności, oraz na znakach wodnych o większej pojemności szum Gaussa.

Tab. 3. BER[%] dla różnych pojemności znaku wodnego oraz szumów

	16 bitów	64 bity	256 bitów	1024 bity
Bez szumu	0	0	0,05	4,58
Szum gaussa	0,01	0,03	1,19	11,19
Kompresja Jpeg [6]	0,12	0,09	0,95	9,11
Rozmycie	0	0	0,06	4,73
Jasność	0,03	0,07	0,55	5,82
Kontrast	0	0	0,05	4,6
Saturacja	0	0	0,05	4,59
Wycięcie	0	0,01	0,34	7,68
Wszystkie	1,11	1,52	7,67	21,28

W przypadku kluczy dłuższych niż 16 bitów, wyniki były mniej satysfakcjonujące. Im dłuższy klucz, tym większe możliwości w stosowaniu dodatkowych kodów nadmiarowych, ale również większa widoczność.

4. PROCES TRENINGOWY

Proces trenowania polegał na nauczeniu modelu, który miał za zadanie generować obrazy z zachowaniem wysokiej pojemności, transparentności i odporności. W tym celu zastosowano kilka modeli o różnych pojemnościach, a każdy z nich był trenowany na tym samym zbiorze danych, wybranym z bazy Hollywood2 [5], która zawierała krótkie klipy video ze scenami z popularnych filmów. Przed rozpoczęciem treningu, filmy były dzielone na klatki i kadrowane na odpowiednie fragmenty o wymiarach 256x256.

Dodatkowo, obrazy te były poddawane dodatkowej augmentacji, czyli modyfikacji w celu zwiększenia zróżnicowania zbioru treningowego. W ramach znaku wodnego w każdej iteracji generowano losowy ciąg bitów, co pozwalało na śledzenie pochodzenia obrazów.

Proces trenowania trwał pięć epok i obejmował 210 000 próbek. W ramach treningu przyjęto trzy funkcje kosztu. Pierwszą z nich był błąd adwersarza (GAN), który miał za zadanie nauczyć modelu generować realistyczne obrazy, które trudno odróżnić od prawdziwych zdjęć. Drugą funkcją kosztu była błąd rekonstrukcji obrazu nośnego (LPIPS [8]), który miał na celu zachowanie podobieństwa do obrazów oryginalnych. Trzecią funkcją kosztu był błąd rekonstrukcji znaku wodnego (BCE), który miał na celu odtworzenie ciągu bitów wykorzystanego jako znak wodny.

Wagi poszczególnych funkcji kosztu były następujące: waga dla błędu adwersarza wynosiła 1, waga dla błędu rekonstrukcji obrazu nośnego wynosiła 2, a waga dla błędu rekonstrukcji znaku wodnego wynosiła 10.

Wagi te określały, jak bardzo dana funkcja kosztu wpłynie na cały proces trenowania. Dzięki odpowiedniemu doborowi wag możliwe było skupienie się na wybranych celach treningowych, na przykład na uzyskaniu bardziej realistycznych obrazów czy też na zachowaniu podobieństwa do oryginalnych zdjęć.

W trakcie pierwszej epoki, ze względu na wysoką złożoność zagadnienia enkodowania i dekodowania, funkcja kosztu dla rekonstrukcji obrazu nośnego została wyłączona. Pozwoliło to na bardziej efektywne nauczenie modelu w początkowych etapach treningu.

5. PODSUMOWANIE

Podsumowując, wyniki eksperymentów pokazują, że algorytm oparty na modelach U-net i ganach może być skuteczny w ukrywaniu danych w obrazach, ale wyniki zależą od długości klucza i rodzaju szumu. W przypadku krótkich kluczy wyniki były bardzo dobre, jednak dla dłuższych kluczy nadal istnieje ryzyko utraty danych podczas procesu dekodowania, szczególnie w przypadku intensywnego szumu.

LITERATURA

[1] Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, i Yoshua Bengio. 2014. „Generative Adversarial Networks”. arXiv. https://doi.org/10.48550/arXiv.1406.2661.

[2] Li, Lingzhi, Jianmin Bao, Hao Yang, Dong Chen, i Fang Wen. 2020. „FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping”. arXiv. https://doi.org/10.48550/arXiv.1912.13457.

[3] Liu, Bingchen, Yizhe Zhu, Kunpeng Song, i Ahmed Elgammal. 2021. „Towards Faster and Stabilized GAN Training for High-fidelity Few-shot Image Synthesis”. arXiv. https://doi.org/10.48550/arXiv.2101.04775.

[4] Lomnitz, Michael R. (2019) 2023. „DiffJPEG: A PyTorch implementation”. Python. https://github.com/mlomnitz/DiffJPEG.

[5] Marszalek, Marcin, Ivan Laptev, i Cordelia Schmid. 2009. „Actions in context”. W *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2929–36. https://doi.org/10.1109/CVPR.2009.5206557.

[6] Morkel, Tayana, Jan Eloff, i Martin Olivier. 2005. „An overview of image steganography”. W , 1–11.

[7] Ronneberger, Olaf, Philipp Fischer, i Thomas Brox. 2015. „U-Net: Convolutional Networks for Biomedical Image Segmentation”. W *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, zredagowane przez Nassir Navab, Joachim Hornegger, William M. Wells, i Alejandro F. Frangi, 234–41. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-24574-4_28.

[8] Zhang, Richard, Phillip Isola, Alexei A. Efros, Eli Shechtman, i Oliver Wang. 2018. „The Unreasonable

Effectiveness of Deep Features as a Perceptual Metric". arXiv.
https://doi.org/10.48550/arXiv.1801.03924.

WPŁYW CIAŁA LUDZKIEGO NA PARAMETRY ELEKTRYCZNE ANTENY TEKSTRONICZNEJ PRZEZNACZONEJ DO PRACY W SYSTEMIE 5G

INFLUENCE OF THE HUMAN BODY ON THE ELECTRICAL PARAMETERS OF THE WEARABLE ANTENNA DESIGNED TO WORK IN THE 5G SYSTEM

Rafał Przesmycki[1]; Marek Bugaj[2]

[1] Wojskowa Akademia Techniczna, ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warszawa, rafal.przesmycki@wat.edu.pl
[2] Wojskowa Akademia Techniczna, ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warszawa, marek.bugaj@wat.edu.pl

DOI: 10.15199/59.2023.4.71

Streszczenie: **W artykule przedstawiono antenę o częstotliwości środkowej 3,6 GHz. Opracowana antena ma zwartą konstrukcję o wymiarach (50 x 56 x 1) mm. Jako podłoże do konstrukcji anteny użyto materiału tekstylnego typu jeans o współczynniku dielektrycznym 1,76 i grubości h=1,0 mm. W artykule przedstawiono analizę wyników symulacji parametrów elektrycznych i charakterystyk promieniowania proponowanej anteny dla przypadku anteny umieszczonej w wolnej przestrzeni oraz anteny z modelem ciała ludzkiego.**

Abstract: **The article presents an antenna with a center frequency of 3.6 GHz. The developed antenna has a compact structure with dimensions (50 x 56 x 1) mm. As a substrate for the construction of the antenna, a jeans-type textile material with a dielectric coefficient of 1.76 and a thickness of h=1.0 mm was used. The article presents an analysis of the simulation results of electrical parameters and radiation characteristics of the proposed antenna for the case of an antenna placed in free space and an antenna with a model of the human body.**

Słowa kluczowe: **antena, antena mikropaskowa, antena tekstroniczna, system 5G.**

Keywords: **antenna, microstrip antenna, wearable antenna, 5G wireless system.**

1. WSTĘP

W dzisiejszych czasach coraz większą rolę w naszym życiu odgrywają sieci bezprzewodowe. Jest to spowodowane między innymi tym, że przesyłamy coraz więcej informacji drogą bezprzewodową. W związku z tym połączenia takie muszą charakteryzować się coraz większą prędkością przesyłania danych oraz jak najmniejszym opóźnieniem. Obecnie wiele państw jest już na etapie wprowadzania systemów telefonii komórkowej piątej generacji – 5G. Systemy te będą pracować na innych zakresach częstotliwości niż obecnie stosowane systemy czwartej generacji, będą to coraz wyższe częstotliwości, w związku z czym będą wymagały innej architektury. W skład takiej architektury wchodzą między innymi anteny nadawcze i odbiorcze, którym poświęcono uwagę w niniejszym artykule.

Drugim ważnym trendem rozwijanym w inżynierii jest miniaturyzacja urządzeń. Miniaturyzacja jest widoczna w niemal każdej dziedzinie techniki: mechanice, optyce, czy elektronice. Analizując miniaturyzację w dziedzinie elektroniki należy wspomnieć o antenach wykonywanych techniką mikropaskową lub z wykorzystaniem tekstyliów, które charakteryzują się między innymi małą wagą oraz objętością. Łącząc to z faktem, że wraz ze wzrostem częstotliwości maleją wymiary elementów promieniujących w antenie, można stwierdzić, że miniaturyzacja anten jest nieunikniona.

Anteny przystosowane do systemów piątej generacji pozwolą rozwinąć wiele dziedzin naszego życia. Coraz bardziej popularne stają się tak zwane inteligentne domy, których celem jest dołączenie do sieci jak najwięcej urządzeń wykorzystywanych w naszym gospodarstwie domowym. Inne dziedziny życia, w których systemy 5G będą odgrywały kluczową rolę to między innymi przemysł, transport miejski czy medycyna (kontrola parametrów życiowych). W medycynie powszechnie zaczynają być stosowane anteny inteligentne, które pomagają w wykonywaniu codziennych czynności. Dzięki ich małej wadze nie przeszkadzają sportowcom w czasie treningów i zawodów, a jednocześnie monitorują i zapisują ich wyniki co pozwala na ulepszenie metod treningowych. Osoby niepełnosprawne również mogą zyskać na tej technologii. Przykładem jest prototyp rękawicy dla osób z dysfunkcją mowy, która zamienia język migowy na mówiony. Istnieje bardzo dużo możliwości zastosowania systemów tekstronicznych, jednak najszybciej rozwijającą się dziedziną gdzie stosuje się te rozwiązania jest odzież ochronna (i nie tylko) dla służb mundurowych [1][7].

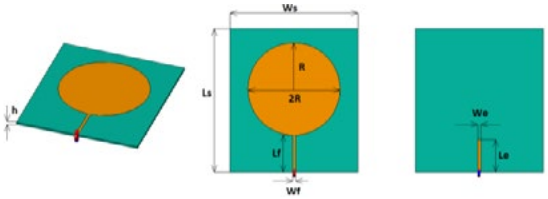
2. MODEL ANTENY TEKSTRONICZNEJ

W chwili obecnej wyróżniamy trzy zakresy częstotliwości zarezerwowane dla omawianej sieci 5G. Pierwsze zakresy częstotliwości jakie będą stosowane w sieciach 5G to: 600 - 700 MHz, 3 - 4 GHz oraz powyżej 26 GHz (26 - 28 GHz i 38 - 42 GHz) [6][4].

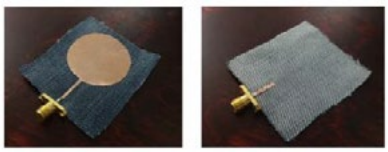
Ze względu na planowane zakresy częstotliwości pracy systemu 5G w Polsce, głównym założeniem dla projektowanej anteny tekstronicznej działającej na jednym z pasm częstotliwości, które ma być wykorzystywane w systemach piątej generacji, jest częstotliwość pracy w zakresie 3,4 – 3,8 GHz. Oprócz pasma częstotliwości innym ważnym wymaganiem stawianym dla projektowanej anteny są wymiary anteny, które nie powinny być większe niż 60 x 60 mm oraz rodzaj materiału, z którego wykonana ma być antena przeznaczona do umiejscowienia jej na jednym z elementów odzieży człowieka. Na tej podstawie wybrano materiał, z którego wykonywana

jest powszechnie odzież użytkowana przez człowieka na co dzień. Materiał ten to jeans. Na bazie analizowanej literatury oraz przeprowadzonych pomiarów parametrów elektrycznych z wykorzystaniem systemu SPEAG DAK-TL-P dla materiału typu jeans przyjęto dla projektu anteny następujące wartości jego parametrów: przenikalność elektryczna ε_r = 1,76, tangens kąta stratności tan(δ) = 0,04, grubość materiału 1 mm.

Dla założonych parametrów na antenę dokonano opracowania modelu w oprogramowaniu do projektowania anten mikropaskowych FEKO firmy Altair. W oprogramowaniu tym można wyznaczyć również parametry elektryczne i charakterystyki promieniowania anteny. W oprogramowaniu tym dokonano procesu optymalizacji pod kątem rozmiaru i częstotliwości pracy opracowywanej anteny. Wykonano proces obliczeniowy z wykorzystaniem promiennika w kształcie koła z zachowaniem parametrów podłoża typu jeans oraz podstawowych założeń na częstotliwość rezonansową i pasmo pracy. Proces wyznaczania parametrów anteny mikropaskowej z promiennikiem kołowym jest analogiczny jak w przypadku anten mikropaskowych z promiennikiem prostokątnym. W wyniku przeprowadzonego procesu optymalizacji ostatecznie otrzymano model anteny przedstawiony na rys. 1, którego wymiary przedstawiono w Tabeli 1 a wygląd wykonanej anteny na rys. 2.



Rys. 1. Finalna wersja anteny tekstronicznej – widok z przodu i widok z tyłu



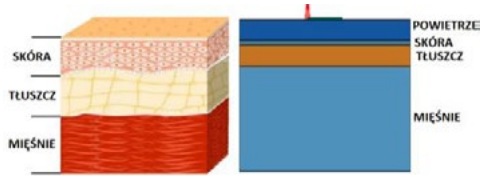
Rys. 2. Wygląd wykonanej anteny tekstronicznej – widok z przodu i widok z tyłu

Tab. 1. Wymiary finalnego projektu anteny

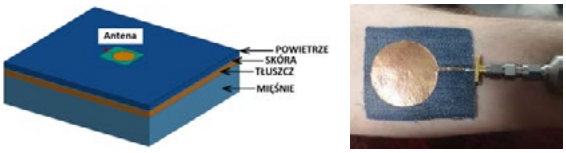
Element anteny	Symbol	Wymiar [mm]
Szerokość podłoża	W_s	50,00
Długość podłoża	L_s	56,00
Szerokość ekranu	W_e	1,14
Długość ekranu	L_e	12,77
Promień łaty	R	18,00
Grubość podłoża	h	1,00
Przenikalność podłoża	ε_r	1,76
Szerokość linii zasilającej	W_f	1,00
Długość linii zasilającej	L_f	19,00

3. MODEL CIAŁA LUDZKIEGO

W celu przedstawienia wpływu ciała ludzkiego na parametry elektryczne anteny tekstronicznej pracującej w systemie 5G opracowano model ciała ludzkiego w oprogramowaniu FEKO. Opracowany model ciała ludzkiego składa się z kilku warstw o określonych parametrach elektrycznych mających wpływ na parametry proponowanej anteny (rys. 2). Przyjęty model ciała ludzkiego składa się z warstwy skóry, tłuszczu i mięśni.



Rys. 3. Wygląd warstw w uproszczonym modelu, z których składa się ciało ludzkie



Rys. 4. Zaprojektowana antena wraz modelem ciała ludzkiego - widok z góry

Pierwszą warstwą oddzielającą antenę od człowieka jest powietrze. Ten ośrodek również wpływa na charakterystykę pracy anteny, dlatego został dodany w środowisku FEKO. Posiada on wymiary 200 x 200 mm, grubość 5 mm oraz przenikalność elektryczną ε_r = 1,001. Kolejną warstwą wprowadzoną, a zarazem pierwszą wchodzącą w skład warstw ciała ludzkiego jest skóra. Jej przenikalność elektryczna oraz tangens kąta strat zostały obliczone poprzez aplikację na stronie internetowej [2] dla częstotliwości 3,6 GHz, a następnie wprowadzone do programu symulacyjnego. Jeśli chodzi o wysokość i szerokość to są one identyczne jak w przypadku warstwy powietrza, jednak grubość wynosi 2 mm, a przenikalność elektryczna ε_r = 36,92 [2][5][8][3]. Kolejnym ośrodkiem, przez który fala elektromagnetyczna przenika w głąb ciała ludzkiego jest tłuszcz. W zależności od osoby może mieć on różną grubość. Na potrzeby symulacji przyjęto, że ten parametr będzie wynosił 5 mm, a przenikalność elektryczna ε_r = 5,16 [2]. Tak samo jak w przypadku poprzedniego ośrodka jego stała dielektryczna została wyliczona, a następnie wprowadzona do oprogramowania FEKO [2][5][8][3]. Ostatnim uwzględnionym ośrodkiem są ludzkie mięśnie. Ich struktura jest najbardziej zbita przez co grubość również jest większa niż dla pozostałych warstw. Przenikalność elektryczna dla częstotliwości 3,6 GHz wynosi ε_r =2.6458 [2][5][8][3]. Wymiary nie uległy zmianie, oprócz wspomnianej grubości, która dla tego ośrodka wynosi 30 mm. Wygląd warstw w uproszczonym modelu, z których składa się ciało ludzkie oraz proponowany model anteny wraz z ciałem ludzkim zostały przedstawione na rys. 3 i rys. 4, a szczegółowe zestawienie wymiarów poszczególnych warstw zostały przedstawione w Tabeli 2.

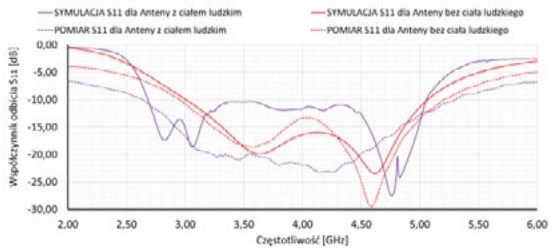
Tab. 2. Wymiary modelu ciała ludzkiego

Warstwa	Wymiary [mm]
Szerokość modelu ciała ludzkiego	200
Długość modelu ciała ludzkiego	200
Grubość powietrza	5
Grubość skóry	2
Grubość tłuszczu	5
Grubość mięśni	30

4. WPŁYW CIAŁA LUDZKIEGO NA PARAMETRY ELEKTRYCZNE ANTENY

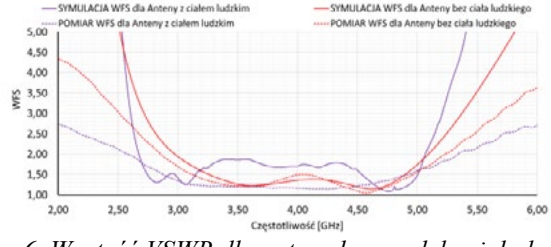
Zadaniem anten tekstronicznych jest przesyłanie i odbieranie danych w zależności od systemów nadawczo odbiorczych, dla których są projektowane. Często znajdują się bezpośrednio w ubraniu osoby, która korzysta z danego systemu nadawczo odbiorczego. Pracują one więc bezpośrednio przy ciele ludzkim, dlatego proponowany model anteny został umieszczony w oprogramowaniu FEKO w pobliżu modelu ciała ludzkiego. W środowisku symulacyjnym zostały więc dodane warstwy mające na celu zobrazowanie jego wpływu na parametry zaprojektowanej konstrukcji. Dla tak zmodyfikowanego modelu przeprowadzono symulacje parametrów anteny w celu pokazania wpływu ciała ludzkiego na parametry elektryczne anteny tekstronicznej pracującej w systemie 5G.

Na rysunku 5 przedstawiono wartości współczynnika odbicia S_{11} dla proponowanej anteny na podłożu typu jeans pracującej w systemie 5G (w wolnej przestrzeni) oraz dla tej samej anteny wraz modelem ciała ludzkiego. W porównaniu do wyników symulacji bez uwzględnienia ciała ludzkiego, możemy zauważyć polepszenie wartości tego parametru, a więc lepsze dopasowanie impedancyjne w szerszym paśmie pracy. Struktura ciała ludzkiego częściowo pochłonęła ilość energii, którą może wypromieniować proponowana antena 5G na podłożu typu jeans.



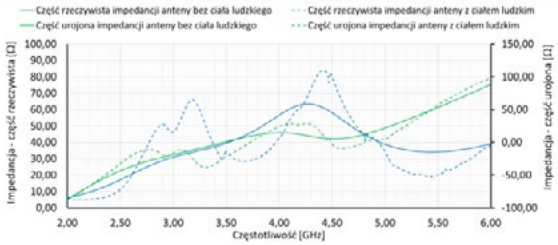
Rys. 5. Wartość S_{11} dla anteny bez modelu ciała ludzkiego oraz anteny wraz modelem ciała ludzkiego

Na rys. 6 przedstawiono wyniki parametru VSWR dla zaprojektowanej anteny pracującej w systemach 5G oraz dla tej samej anteny wraz modelem ciała ludzkiego. Z racji poszerzenia i przesunięcia pasma pracy, najlepszy wynik wartości VSWR występuje na innej częstotliwości niż dla modelu bez modelu ciała ludzkiego, ale nie ma to wpływu na pracę anteny w paśmie dla systemu 5G.



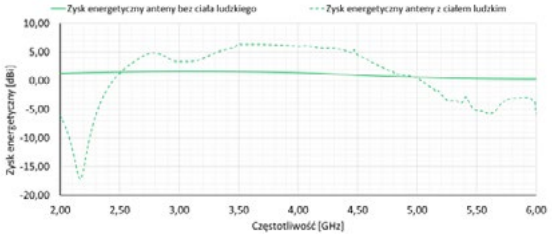
Rys. 6. Wartość VSWR dla anteny bez modelu ciała ludzkiego oraz anteny wraz modelem ciała ludzkiego

Impedancja wejściowa składa się z części rzeczywistej i urojonej. Porównanie tego parametru zostało przedstawione na rys. 7. Jak możemy zaobserwować dla części odpowiadającej za rezystancję promieniowania oraz rezystancję strat, charakterystyka jest dużo bardziej równomierna dla modelu bez uwzględnionego ciała ludzkiego. Antena jest lepiej dopasowana niż antena pracująca z modelem ciała ludzkiego i w całym zakresie pracy nie przekracza 60 Ω. Jeśli chodzi o antenę z modelem ciała ludzkiego wartość impedancji wejściowej przyjmuje wyniki w okolicy wartości 50 Ω. Z tego względu można powiedzieć, że ciało ludzkie znacząco wpływa na dopasowanie impedancyjne anteny, co może być niekiedy przeszkodą w stosowaniu takich anten przy ciele ludzkim bez konieczności stosowania dodatkowych elementów odpowiadających za dopasowanie impedancyjne anteny.

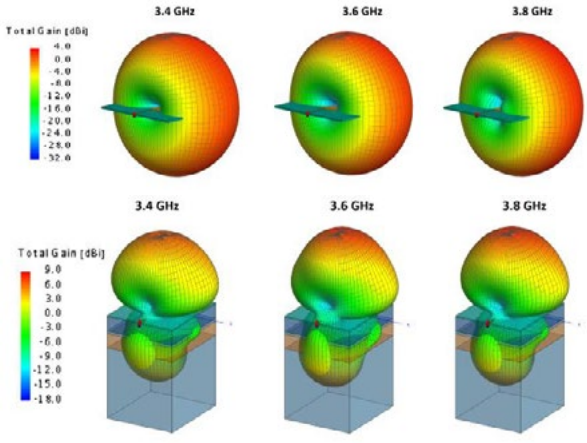


Rys. 7. Wartość impedancji wejściowej (części rzeczywistej i części urojonej) dla anteny bez modelu ciała ludzkiego oraz anteny wraz modelem ciała ludzkiego

Na rys. 8 przedstawiono wyniki zysku energetycznego dla zaprojektowanej anteny w wolnej przestrzeni oraz dla tej samej anteny wraz modelem ciała ludzkiego. Dla proponowanej anteny antena z uwzględnieniem ciała ludzkiego zysk energetyczny zwiększył się do wartości około 6,2 dBi.



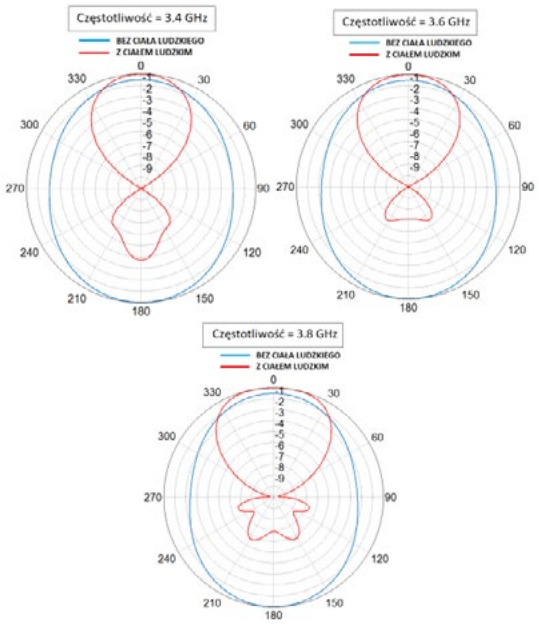
Rys. 8. Zysk kierunkowy zaprojektowanej anteny bez i z uwzględnieniem ciała ludzkiego



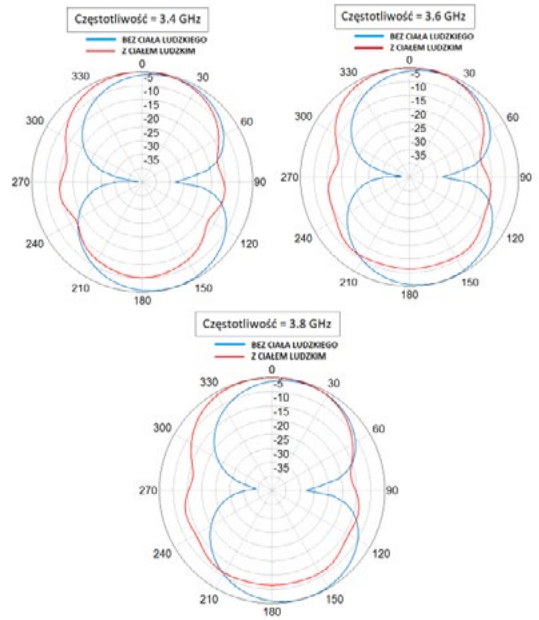
Rys. 9. Charakterystyka promieniowania 3D zaprojektowanej anteny z uwzględnionym modelem ciała ludzkiego dla częstotliwości 3.4 GHz, 3.6 GHz, 3.8 GHz

Ostatnim porównywanym parametrem proponowanej anteny jest charakterystyka promieniowania anteny w wolnej przestrzeni i anteny z modelem ciała ludzkiego,

która została przedstawiona na rys. 9 dla częstotliwości 3.4 GHz, 3.6 GHz oraz 3.8 GHz.



Rys. 10. Unormowane charakterystyki promieniowania anteny bez i z uwzględnieniem ciała ludzkiego dla częstotliwości 3,4 GHz, 3,6 GHz, 3,8 GHz dla polaryzacji pionowej (w płaszczyźnie E)



Rys. 11. Unormowane charakterystyki promieniowania anteny bez i z uwzględnieniem ciała ludzkiego dla częstotliwości 3,4 GHz, 3,6 GHz, 3,8 GHz dla polaryzacji poziomej (w płaszczyźnie H)

Na rys. 10 przedstawiono unormowane charakterystyki promieniowania proponowanej anteny dla częstotliwości 3,4 GHz, 3,6 GHz oraz 3,8 GHz bez i z uwzględnieniem ciała ludzkiego w układzie współrzędnych biegunowych dla polaryzacji pionowej. Na rys. 11 przedstawiono unormowane charakterystyki promieniowania proponowanej anteny dla częstotliwości 3,4 GHz, 3,6 GHz oraz 3,8 GHz bez i z uwzględnieniem ciała ludzkiego w układzie współrzędnych biegunowych dla polaryzacji poziomej.

5. WNIOSKI

Zaprojektowana antena tekstroniczna może być wykorzystywana w codziennym życiu przez zwykłego użytkownika, oraz przez służby mundurowe w czasie rutynowych czynności. Konstrukcja anteny z racji swoich wymiarów nie powinna wpływać negatywnie na zakres ruchów oraz oddziaływać niekorzystnie na zdrowie i życie ludzkie. Zaprojektowana antena może służyć do komunikacji w gęsto zaludnionych miastach, gdzie znajduje się wiele przeszkód terenowych, które w przypadku sieci 4G utrudniają przesył informacji. Materiałem użytym na podłoże anteny jest materiał tekstylny typu jeans. Jego powszechność, niska cena, uniwersalność oraz stosunkowo duża wytrzymałość przyczyniły się do wyboru tej tkaniny, jako głównego elementu konstrukcji anteny.

Opisana w artykule antena na podłożu typu jeans przeznaczona do pracy w systemie 5G bez uwzględnienia ciała ludzkiego charakteryzuje się częstotliwością środkową równą 3.63 GHz, niskim współczynnikiem odbicia -23.41 dB, wartością zysku energetycznego 1.54 dBi, szerokim pasmem pracy 2.121 GHz (52.36%) oraz wysoką sprawnością energetyczną. Proponowana antena z uwzględnieniem ciała ludzkiego ma częstotliwość środkową równą 3.87 GHz, niski współczynnik odbicia -27,54 dB, wartość zysku energetycznego 6.34 dBi, szerokie pasmo pracy 2.371 GHz (61.26%) oraz wysoką sprawność energetyczną. Poprawa parametrów anteny wynika z faktu, że ciało człowieka stanowi niejako ekran dla anteny i powoduje skupienie energii w jednym kierunku.

Praca została sfinansowana przez Wojskową Akademię Techniczną w ramach projektu nr UGB-22-859/2023.

LITERATURA

[1] Ewa Skrzetuska, *„Trendy rozwojowe w tekstronice - Rozwiązania tekstroniczne”*, Przegląd Elektrotechniczny nr 4/2014, str. 34-40.
[2] http://niremf.ifac.cnr.it/tissprop/htmlclie/htmlclie.php
[3] https://pl.dreamstime.com/struktura-ludzka-sk%C3%B3ra-image101128233
[4] https://www.gov.pl/web/5g/czestotliwosci
[5] Ibanez-Labiano, I.; Alomainy, A. *Dielectric Characterization of Non-Conductive Fabrics for Temperature Sensing through Resonating Antenna Structures. Materials* 2020, 13, 1271. https://doi.org/10.3390/ma13061271
[6] Ministerstwo Cyfryzacji, Instytut Łączności, Collegium Medicum UJ, *„Przewodnik po 5G”*, 2019.
[7] Sylwia Walczak, *„Inteligentne tekstylia - międzynarodowe innowacje w tekstronice”*, Acta Innovations nr 3/2012, str. 103-122.
[8] W. Farooq, M. Ur-Rehman, Q. H. Abbasi and K. Qaraqe, *"A circular patch frequency reconfigurable antenna for wearable applications"* 2015 IEEE 11th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), Abu Dhabi, 2015, pp. 103-106, doi: 10.1109/WiMOB.2015.7347947.

POMIAR CHARAKTERYSTYK PROMIENIOWANIA ANTEN Z WYKORZYSTANIEM IMPULSÓW ELEKTROMAGNETYCZNYCH HPM

MEASUREMENT OF ANTENNA RADIATION PATTERNS USING HPM ELECTROMAGNETIC PULSE

Rafał Przesmycki[1]; Marek Bugaj[2]

[1] Wojskowa Akademia Techniczna, ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warszawa, rafal.przesmycki@wat.edu.pl
[2] Wojskowa Akademia Techniczna, ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warszawa, marek.bugaj@wat.edu.pl

DOI: 10.15199/59.2023.4.72

Streszczenie: **W artykule przedstawiono metodę pomiaru charakterystyk promieniowania anten z wykorzystaniem impulsów elektromagnetycznych. Jako źródło impulsów elektromagnetycznych wykorzystano generator HPM DS110 firmy DIEHL. W celu weryfikacji poprawności zaproponowanej metody pomiarowej dokonano pomiaru charakterystyk promieniowania dwóch anten z wykorzystaniem zaproponowanej metody oraz metody klasycznej. Dokonano porównania uzyskanych wyników. Wskazano zalety i ograniczenia zaproponowanej metody pomiaru charakterystyk promieniowania anten.**

Abstract: **The article presents a method of measuring antenna radiation patterns using electromagnetic pulses. DIEHL HPM DS 110 generator was used as the source of electromagnetic pulses. In order to verify the correctness of the proposed measurement method, the radiation patterns of two antennas were measured using the proposed method and the classical method. The obtained results were compared. The advantages and limitations of the proposed method of measuring antenna radiation patterns were indicated.**

Słowa kluczowe: **charakterystyka promieniowania, antena, impuls elektromagnetyczny, HPM.**

Keywords: **radiation pattern, antenna, electromagnetic pulse, HPM.**

1. WSTĘP

W obecnych czasach automatyzacja różnych procesów, badań, czy pomiarów jest powszechnie wykorzystywana. Wprowadza się ją by nie musieć wykonywać ręcznie żmudnych, długich i powtarzalnych czynności. Dodatkowo, pozytywnym aspektem automatyzacji takich zabiegów jest znaczne poprawienie ich dokładności. Przykładem takiego procesu może być pomiar charakterystyki promieniowania anteny. Pomiary są dokonywane dla zakresu 360 stopni i często zajmują dużo czasu. Bardzo ważnym aspektem jest tutaj dokładność. Człowiek nie jest w stanie za każdym razem obrócić anteny o dokładnie 1 stopień lub mniej. Automatyzacja tego pomiaru eliminuje ten problem, a dodatkowo wymaga od operatora jedynie wprowadzenia kilku niezbędnych parametrów i nadzorowania wykonywanych pomiarów, bez konieczności wykonywania powtarzalnych czynności. Automatyzacja nie tylko poprawia dokładność badań, ale również znacząco poprawia jakość pracy operatora, pozwalając skupić mu się na analizie przeprowadzanych badań, a nie na czynnościach wykonywanych przez automat. Dokładne zbadanie charakterystyki promieniowania anteny jest istotnym zagadnieniem z punktu widzenia wielu systemów, np. radiokomunikacyjnych, radiolokacyjnych, telefonii komórkowej, czy Wi-Fi. Odpowiednio zbadana charakterystyka promieniowania pozwala na bardziej efektywne wykorzystanie anten m. in. w wyżej wymienionych systemach. W artykule zaproponowano nową metodę pomiaru charakterystyk promieniowania anten bazującą na wykorzystaniu impulsów HPM (High Power Microwave) [3][7].

2. METODY POMIARU

Charakterystyka promieniowania w sposób graficzny przedstawia rozkład pola elektromagnetycznego na powierzchni kuli będącej strefą promieniowania danej anteny. Środek tej kuli pokrywa się ze środkiem anteny, w związku z czym charakterystyka ta nie zależy od odległości od anteny, ale od kątów ϕ i θ [3][7].

Metody pomiaru charakterystyki promieniowania można podzielić w zależności od właściwości wykonywanego pomiaru. Do najważniejszych metod pomiaru charakterystyk promieniowania anten możemy zaliczyć:

- metody pomiaru charakterystyk promieniowania anten w strefie bliskiej,
- metody pomiaru charakterystyk promieniowania anten w strefie dalekiej.

W celu wyznaczenia charakterystyki promieniowania anteny w polu dalekim dokonując pomiaru jednej składowej pola można uzyskać pełną informację o drugiej składowej oraz gęstości mocy promieniowania. Natomiast w polu bliskim należy oddzielnie zmierzyć składową elektryczną i magnetyczną pola elektromagnetycznego, co jest bardziej czasochłonne i wymaga specjalistycznej aparatury pomiarowej.

W metodach pomiaru charakterystyk promieniowania anten w strefie bliskiej realizuje się pomiary w strefie bliskiej pola elektromagnetycznego oraz wykorzystuje się modele matematyczne, bazujące na potencjałach wektorowych. Pomiary wykonuje się z bliskiej odległości skanując zadaną powierzchnię anteny z wykorzystaniem odpowiednich skanerów pola bliskiego. Dzięki pomiarom rozkładu amplitudy promieniowania, fazy oraz polaryzacji możliwe jest obliczenie rozkładu pola w polu dalekim z wykorzystaniem metod syntezy charakterystyk promieniowania, które są wybierane w zależności od kształtu

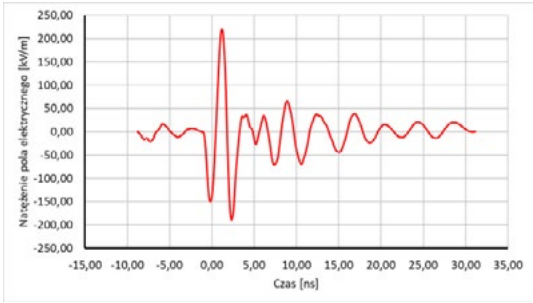
wiązki głównej i listków bocznych. Wartości pola elektromagnetycznego mierzone są w dyskretnych regularnie rozmieszczonych punktach w strefie bliskiej, a następnie transformowane są do strefy dalekiej, co pozwala wyznaczyć przestrzenną charakterystykę promieniowania.

W metodach pomiaru charakterystyk promieniowania anten w strefie dalekiej stosowana jest metoda polegająca na pomiarze odbieranej mocy lub napięcia w antenie odbiorczej (antenie badanej) wygenerowanej przez pomocniczą antenę nadawczą umiejscowioną w odległości spełniającej kryterium strefy dalekiej. Podczas pomiaru charakterystyki promieniowania obraca się anteną badaną w polu fali elektromagnetycznej wytworzonej przez antenę pomocniczą i rejestruje się punkt po punkcie wartość sygnału na wyjściu anteny badanej dla danego ustawienia kątowego anteny i dla każdej częstotliwości oddzielnie. Zwykle częstotliwość pracy dla badanej anteny jest podawana jako parametr o wartości stałej. Więc charakterystyka promieniowania jest mierzona dla tej określonej częstotliwości (lub kilku częstotliwości). Niepraktyczne jest mierzenie charakterystyki promieniowania w sposób ciągły. W przypadku wyznaczenia charakterystyk promieniowania anteny dla wielu częstotliwości z pasma pracy staje się to bardzo czasochłonne. Z tego względu w artykule zaproponowano metodę pomiaru charakterystyki promieniowania anteny z wykorzystaniem impulsów HPM, dzięki której możliwe jest uzyskanie w jednym pomiarze kilkudziesięciu, a nawet kilkuset charakterystyk promieniowania.

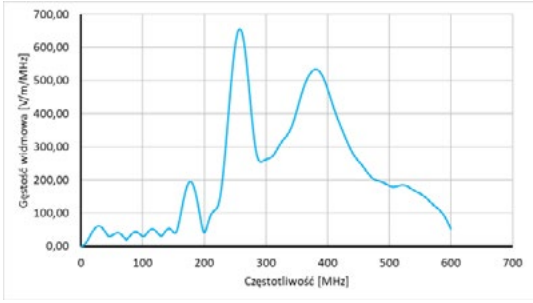
Metoda pomiaru charakterystyk promieniowania anteny z wykorzystaniem impulsów HPM polega na pomiarze w dziedzinie czasu impulsu elektromagnetycznego HPM z wykorzystaniem anteny odbiorczej (anteny badanej) wygenerowanego przez wysokomocowy generator impulsów HPM DS110 umiejscowiony w odległości spełniającej kryterium strefy dalekiej. Podczas pomiaru charakterystyk promieniowania anteny obraca się anteną badaną i rejestruje się impulsy HPM w dziedzinie czasu na wyjściu anteny badanej dla danego ustawienia kątowego anteny. Dzięki pomiarom impulsów w dziedzinie czasu dla wszystkich ustawień kątowych anteny badanej możliwe jest obliczenie gęstości widmowej dla wszystkich zarejestrowanych impulsów HPM z wykorzystaniem szybkiej transformaty Fouriera FFT uwzględniającej parametry całego systemu pomiarowego. Na podstawie uzyskanych wyników otrzymujemy charakterystyki promieniowania badanej anteny dla wszystkich częstotliwości wchodzących w skład wyznaczonych gęstości widmowych zarejestrowanego impulsu HPM.

3. CHARAKTERYSTYKA IMPULSU HPM

Rozwój technologiczny w zakresie mikrofal dużej mocy sprawił, że możliwa jest produkcja i większa dostępność bardziej wydajnych źródeł mikrofalowych. Jednocześnie, możliwa jest generacja pól o stosunkowo dużych natężeniach z wykorzystaniem systemów kompaktowych, które mogą mieścić się, np. w walizce. Impuls elektromagnetyczny wielkiej mocy określany jako HPM jest to impuls o bardzo krótkim czasie trwania (rzędu nanosekund) i o wielkiej mocy (rzędu kilkudziesięciu gigawatów). Na ogół, ilość rodzajów kształtów fali HPM jest ograniczona do czterech. Parametry kształtu fali, takie jak częstotliwość, moc, itp. zależą od rodzaju źródła. Do podstawowych kształtów fali przebiegów HPM zaliczamy: falę ciągłą (CW – Continuous Wave), wąskopasmową (NB – Narrow Band), sinusoidalnie tłumioną (DS – Damped Sinusoid) oraz szerokopasmową (UWB – Ultra Wide Band). W ramach opracowanej metody pomiaru charakterystyk promieniowania anten z wykorzystaniem impulsu HPM wykorzystano kształt fali DS. Impuls HPM-DS (ang. High Power Microwave Damped Sinusoid) charakteryzuje się częstotliwością środkową oraz dosyć szerokim pasmem częstotliwościowym. Zaletą kształtu fali HPM-DS jest szersze pasmo aniżeli w przypadku fali CW lub NB [1][5]. Na rys. 1 przedstawiono przykładowy zarejestrowany przebieg czasowy impulsu HPM-DS natomiast na rys. 2 wyznaczoną gęstość widmową dla zarejestrowanego impulsu z wykorzystaniem szybkiej transformaty Fouriera FFT.



Rys. 1. Kształt przykładowego zarejestrowanego impulsu HPM-DS w dziedzinie czasu



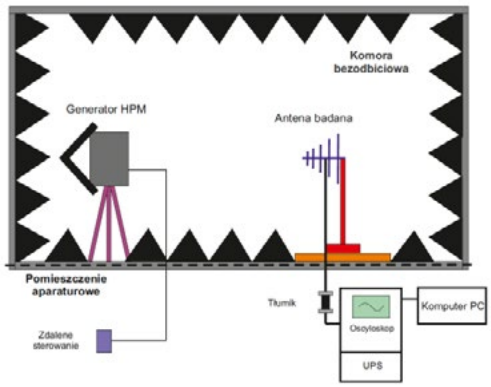
Rys. 2. Gęstość widmowa przykładowego zarejestrowanego impulsu HPM-DS

4. STANOWISKO DO POMIARU CHARAKTERYSTYK PROMIENIOWANIA ANTEN

Ze względu na generowanie wysokomocowych impulsów HPM podczas pomiarów charakterystyk promieniowania stanowisko pomiarowe powinno być umieszczone w komorze bezodbiciowej. Komora bezodbiciowa to pomieszczenie, które jest odpowiednio ekranowane oraz wyposażone w absorbery w.cz.. Ogranicza to w znaczny sposób wpływ zewnętrznych pól elektromagnetycznych na wyniki pomiarów oraz eliminuje fale elektromagnetyczne, które mogły by wydostać się na zewnątrz.

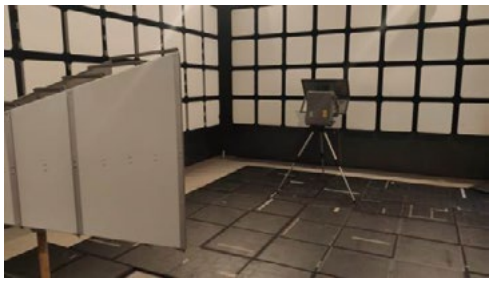
Na rys. 3 zaprezentowano schemat stanowiska pomiarowego do pomiaru charakterystyki promieniowania anten z wykorzystaniem impulsów HPM-DS, natomiast na rys. 4. przedstawiono wygląd zestawionego stanowi-

ska pomiarowego do pomiaru charakterystyk promieniowania anten z wykorzystaniem impulsów HPM. Schemat blokowy stanowiska pomiarowego zawiera elementy umożliwiające automatyczny pomiar charakterystyki promieniowania anteny, za pomocą urządzeń pomiarowych. Sterowanie pomiarami oraz obrotem anteny badanej odbywa się w pomieszczeniu zewnętrznym z wykorzystaniem specjalistycznego oprogramowania. Stanowisko pomiarowe zbudowano na bazie wyposażenia Laboratorium Kompatybilności Elektromagnetycznej WEL WAT. W skład stanowiska pomiarowego wchodzą: antena badana, generator HPM, stół obrotowy, oscyloskop oraz komputer sterujący procesem pomiarowym.



Rys. 3. Schemat blokowy stanowiska pomiarowego do pomiaru charakterystyk promieniowania anten z wykorzystaniem impulsów HPM

Elementem odpowiedzialnym za generację impulsu HPM-DS jest Generator Diehl DS110, który przedstawiono na rys. 5. Jest to kompaktowe urządzenie zamknięte w obudowie, wyglądem przypominające walizkę. Elementami składowymi tego generatora są wysokonapięciowe układy zasilania, generator Marxa 300 kV oraz antena rezonansowa. Wysokonapięciowy układ zasilania dostarcza do generatora Marxa napięcie 50 kV (napięcie symetryczne ± 25 kV), jest on ładowany do momentu ustalonego przez próg przełącznika, który ustala ciśnienie gazu N_2 w generatorze. W momencie zmiany położenia przełącznika do anteny zostaje doprowadzone napięcie o wartości 300 kV, iskrownik znajdujący się w antenie jest izolowany przez ciśnienie gazu N_2 [6][4].



Rys. 4. Widok stanowiska do pomiaru charakterystyk promieniowania anten z wykorzystaniem impulsów HPM

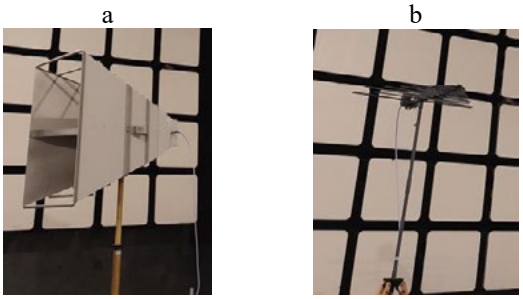
Wyzwalanie impulsów z generatora odbywa się przy pomocy przełącznika zdalnego sterowania połączonego z generatorem przy pomocy światłowodu dla każdego ustawienia kątowego stołu obrotowego z anteną badaną. Oscyloskop odbiera sygnał z anteny badanej i przy pomocy interfejsu ethernetowego przekazuje próbki do komputera pomiarowego, gdzie przy pomocy dedykowanego oprogramowania sygnał jest całkowany dzięki czemu otrzymujemy postać czasową sygnału faktycznie wygenerowanego z generatora oraz realizowana jest szybka transformata Fouriera. Dzięki takiemu procesowi możliwy jest pomiar charakterystyki promieniowania anteny z wykorzystaniem impulsów HPM-DS.

Rys. 5. Widok Generatora Diehl DS 110

5. WYNIKI POMIRÓW

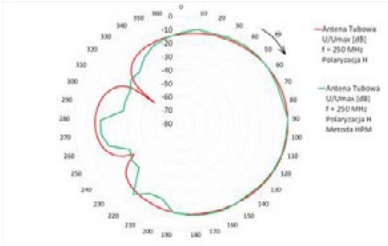
W celu walidacji stanowiska pomiarowego do pomiaru charakterystyk promieniowania anten z wykorzystaniem impulsów HPM-DS przeprowadzono na tym stanowisku pomiary charakterystyk promieniowania dwóch wybranych anten, których wygląd pokazano na rys. 6. Pierwsza antena jest to antena tubowa BBHA9120F pracująca w zakresie częstotliwości od 200 MHz do 2 GHz, natomiast druga antena to antena logoperiodyczna HL223 pracująca w zakresie od 200 MHz do 1300 MHz. Dodatkowo wykonano również dla obydwu wybranych anten pomiary na tradycyjnym stanowisku pomiarowym do badań charakterystyk promieniowania anten w strefie dalekiej dla kilku częstotliwości, które przyjęto jako pomiar odniesienia. Wyniki pomiarów, które zostały zrealizowane na obydwu stanowiskach pomiarowych zostały porównane ze sobą w celu weryfikacji i oceny opracowanego stanowiska pomiarowego.

a b

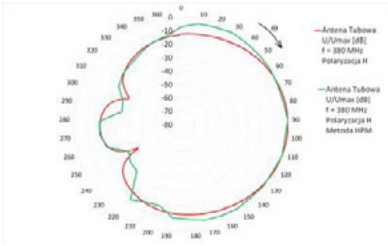


Rys. 6. Wygląd anten badanych: a – antena tubowa BBHA9120F; b – antena logoperiodyczna HL223

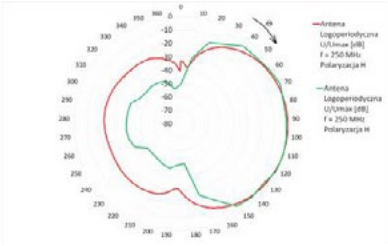
Na rys. 7 i rys. 8 przedstawiono wyniki pomiarów w postaci unormowanych charakterystyk promieniowania anteny tubowej BBHA9120F dla częstotliwości 350 MHz i 400 MHz przy polaryzacji poziomej uzyskane dwoma metodami pomiarowymi. Natomiast na rys. 9 i rys. 10 przedstawiono wyniki pomiarów w postaci unormowanych charakterystyk promieniowania anteny logoperiodycznej HL223 dla częstotliwości 350 MHz i 400 MHz przy polaryzacji poziomej uzyskane dwoma metodami pomiarowymi.



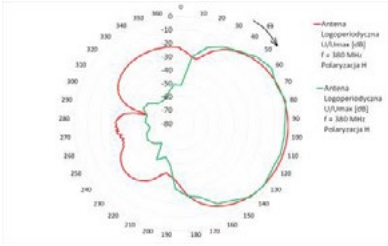
Rys. 7. Unormowane charakterystyki promieniowania anteny BBHA9120F dla częstotliwości 250 MHz



Rys. 8. Unormowane charakterystyki promieniowania anteny BBHA9120F dla częstotliwości 380 MHz



Rys. 9. Unormowane charakterystyki promieniowania anteny HL223 dla częstotliwości 250 MHz



Rys. 10. Unormowane charakterystyki promieniowania anteny HL223 dla częstotliwości 380 MHz

Na podstawie uzyskanych wyników pomiarowych można stwierdzić, że opracowana metoda pomiaru i stanowisko laboratoryjne jest poprawne. Wyznaczając parametry charakterystyk promieniowania anten w postaci szerokości użytecznej i szerokości całkowitej listka głównego można powiedzieć, że wartości te są porównywalne. Na podstawie tego można powiedzieć, że wyniki pomiarowe uzyskane za pomocą metody wykorzystującej impulsy HPM-DS nie odbiegają znacząco od charakterystyk promieniowania badanych anten pomierzonych metodą tradycyjną w strefie dalekiej. Jedynie różnice widać w poziomach listków wstecznych, co świadczy, że charakterystyka promieniowania anteny wyznaczona za pomocą stanowiska tradycyjnego została pomierzona z większą dokładnością pomiaru wynikającą z większego zakresu dynamiki wykorzystywanej aparatury pomiarowej. Dla stanowiska wykorzystującego impulsy HPM-DS największą zaletą jest szybkość pomiaru, która jest nieporównywalna z czasem pomiaru metodą tradycyjną. W opracowanej metodzie po jednym obrocie anteny badanej o 360 stopni uzyskujemy od kilkudziesięciu do kilkuset charakterystyk na różnych częstotliwościach. Wadą opracowanej metody może być wartość generowanego natężenia pola elektromagnetycznego, która osiąga kilkaset kV/m, Z tego względu metody tej nie można stosować dla anten o małej wytrzymałości elektrycznej a raczej dla anten metalowych. Kolejną wadą tej metody jest jej zakres stosowania pod względem częstotliwości. Metodę tą możemy stosować jedynie dla zakresu częstotliwości, który pokrywa wyznaczone widmo impulsu HPM-DS, tj. od kilku MHz do około 600 MHz.

6. WNIOSKI

W artykule przedstawiono opracowaną metodę pomiaru i stanowisko laboratoryjne do automatycznego pomiaru charakterystyk promieniowania anten z wykorzystaniem impulsów HPM-DS.

Na bazie przedstawionych wyników pomiarowych można stwierdzić, że opracowane automatyczne stanowisko pomiarowe charakteryzuje się dużą szybkością pomiaru, która jest nieporównywalna z czasem pomiaru metodą tradycyjną. W opracowanej metodzie po jednym obrocie anteny badanej o 360 stopni uzyskujemy od kilkudziesięciu do kilkuset charakterystyk na różnych częstotliwościach. Taką samą ilość charakterystyk metodą tradycyjną uzyskalibyśmy po wielu godzinach a nawet dniach pomiarowych. Opracowane stanowisko pomiarowe może być wykorzystywane w zakresie częstotliwości do 600 MHz, ponieważ spełnia wymagania procedur badawczych określanych przez laboratorium.

Praca została sfinansowana przez Wojskową Akademię Techniczną w ramach projektu nr UGB-22-859/2023.

LITERATURA

[1] AECTP 250 Ed.2 2011 : *Electrical and electromagnetic environmental conditions*, NATO international staff - defence investment division, 2011.

[2] DIEHL BGT Defence – *Operating and Maintenance Manual for DIEHL DS110.KS2.F.MP1.B (HPM Case System DS110)*, document No BDB8118952.

[3] G. Różański, M. Wnuk, W. Kołosowski, *"Wyznaczanie charakterystyk kierunkowości anten mikrofalowych na podstawie pola w strefie bliskiej"*, Biuletyn WAT, Warszawa 2006

[4] Przesmycki, R.; Bugaj, M. – *Pomiar tłumienności materiałów ekranujących z wykorzystaniem generatora HPM*. Przegląd Elektrotechniczny, R. 97 NR 10/2021, Str. 133 – 139, ISSN 0033-2097,

[5] Rafał Przesmycki – *Classification of the electromagnetic effects of information devices during high power microwave exposing*, IEEE Xplore, Electronic ISBN: 978-1-5386-1211-8, Print on Demand(PoD) ISBN: 978-1-5386-1212-5, Pages 357 – 363,

[6] Rafał Przesmycki, Marian Wnuk – *Susceptibility of IT devices to HPM pulse*, International Journal of Safety and Security Engineering, Vol. 8, No. 2 (2018) 1–11, ISSN 2041-9031, str: 223-233

[7] Standard IEEE 149-1979 „*Procedury testowe dla anten"*

ANALIZA WPŁYWU MIKRO-STRUKTURY PREZENTACJI TREŚCI NA SUBIEKTYWNĄ OCENĘ JAKOŚCI WIDEO I JEJ STABILNOŚĆ

ANALYSIS OF THE INFLUENCE OF CONTENT PRESENTATION MICRO-STRUCTURE ON THE SUBJECTIVE ASSESSMENT OF VIDEO QUALITY AND ITS STABILITY

Tomasz Konaszyński[1]; Dawid Juszka[2]; Mikołaj Leszczuk[3]

[1] Szkoła Doktorska AGH, Kraków, Tomasz.Konaszynski@agh.edu.pl
[2] Instytut Telekomunikacji AGH, Kraków, Dawid.Juszka@agh.edu.pl
[3] Instytut Telekomunikacji AGH, Kraków, Mikolaj.Leszczuk@agh.edu.pl

DOI: 10.15199/59.2023.4.73

Streszczenie: **Przeprowadzono badanie jakości wideo postrzeganej przez użytkownika (Quality of Experience, QoE) i zweryfikowano czy i jak mikro-struktura prezentacji sekwencji wpływa na subiektywną ocenę. W tym celu przeanalizowano wpływ: zmienności/powtarzalności treści, jakości sekwencji poprzedzającej, kolejności sekwencji. Opisano obserwacje dotyczące korelacji QoE z mikro-strukturą prezentacji sekwencji, mogące stanowić podstawę dla hipotez dotyczących zależności oceny QoE od wymienionych czynników, podlegające dalszej pracy badawczej.**

Abstract: **A study of video quality perceived by the user (Quality of Experience, QoE), examining whether and how the order and structure of the video sequence presentation affect the subjective assessment, was conducted. Impact of content variability/repeatability and quality of the preceding sequence or sequence order were analyzed. Observations on the correlation of QoE with the micro-structure of sequence presentation as the basis for hypotheses for the dependence of QoE assessment on the impact factors, were also described.**

Słowa kluczowe: **mikro-struktura treści, Absolute Category Rating, ocena subiektywna, QoE**
Keywords: **content micro-structure, Absolute Category Rating, Quality of Experience (QoE), subjective assessment**

1. WSTĘP

Powszechność usług strumieniowych i potrzeba adaptacji jakości multimediów do potrzeb odbiorcy są czynnikami stymulującymi badania czynników wpływających na percepcję jakości wideo przez odbiorcę. Zagadnienia te wpisują się w obszar Quality of Experience. [7][2][8].

Subiektywna ocena jakości wideo jest wypadkową szeregu czynników wpływających na percepcję oglądającego. Wśród nich w literaturze [10] wymienia się m.in. takie jak: aspekt technologiczny, społeczny, środowiskowy, ludzki czy charakterystykę prezentowanej treści.

Wartym dodatkowej uwagi aspektem jest natomiast potencjalny wpływ **mikro-struktury prezentacji treści**, tj. względnego układu treści wideo, w tym kolejności sekwencji, częstotliwości i krotności wyświetleń, czy też odniesienia do tła – jakości sekwencji wyświetlanych wcześniej.

2. ZAGADNIENIE BADAWCZE

Przedmiotem publikacji jest opis eksperymentu dotyczącego wpływu na ocenę subiektywną czynnika mikro-struktury i układu prezentacji treści wideo, przy neutralności względem czynników społecznych, środowiskowych i ludzkich. Celem autorów było znalezienie odpowiedzi na pytanie badawcze: **Czy na ocenę subiektywną wideo mogą mieć wpływ takie czynniki jak (1) jakość obiektywna danego filmu, (2) krotność wyświetlenia, (3) jakość filmu poprzedzającego.**

3. OPIS EKSPERYMENTU

Projekt eksperymentu. Eksperyment został przeprowadzony zgodnie z zaleceniami Rekomendacji ITU-T P.913 [5]. Celem eksperymentu było pozyskanie ocen subiektywnych wystawionych przez testerów dla zbioru sekwencji PVS (*Processed Video Sequences)* wyświetlanych w odpowiednim porządku i krotności. Testerów podzielono na dwie rozłączne grupy: (1) grupa REGULAR, w której każdy tester obejrzał i ocenił każdą ze 170 sekwencji PVS dokładnie jeden raz; (2) grupa REPEAT, w której każdy tester obejrzał i ocenił każdą z 60 sekwencji PVS trzy razy.

Zadaniem testera było określenie subiektywnie odczuwanej jakości obejrzanej sekwencji (QoE) i przypisanie jej oceny w 5-stopniowej skali Absolute Category Rating (ACR), tj. (5 – excellent, 4 – good, 3 – fair, 2 – poor, 1 – bad), [4]. Tester oceniał jakość po obejrzeniu każdego PVS. Testerzy oceniali sekwencje w tych samych, jednorodnych warunkach technicznych i środowiskowych.

Rezultatem eksperymentu jest zbiór ocen, dla obydwu grup, powiązany z dodatkowymi danymi, takimi jak ID testera, ID filmu, poziom deklarowanej jakości filmu, data i czas rejestracji oceny.

Dane badawcze. W eksperymencie został wykorzystany zestaw sekwencji wideo FullHD wyselekcjonowany spośród sekwencji udostępnionych publicznie w bazach Netflix, CableLabs, SJTU Media Lab oraz Xiph.org Video Test Media. Wybrano 34 unikalne pod względem treści źródłowych sekwencje wideo SRC (*Source Reference Circuit*). Zbiór sekwencji PVS powstał przy użyciu popularnej obiektywnej metryki jakości wideo VMAF (Video

Multimethod Assessment Fusion) [9]. Dla każdej sekwencji SRC wygenerowano pliki w standardzie YUV4MPEG2, utworzone przy poziomach bitrate, zmieniających się w przedziale pomiędzy 100000 b/s a 21000000 b/s.

Każda z sekwencji PVS została oceniona przy użyciu VMAF w zakresie 0–100, gdzie wartość 100 oznacza jakość identyczną z wideo referencyjnym. Dla każdej z wybranych 34 sekwencji SRC do odpowiedniej grupy jakościowej (A-E) została wybrana jedna sekwencja, której obiektywna ocena VMAF znajdowała się najbliżej wartości odpowiednio 90 (Grupa A), 70 (Grupa B), 50 (Grupa C), 30 (Grupa D) i 10 (Grupa E). Zatem dla każdej z 34 sekwencji SRC wybrano 5 sekwencji PVS (reprezentujących pięć poziomów jakości A-E), uzyskując w docelowym zbiorze łącznie 170 sekwencji PVS. Każda sekwencja PVS trwała 10 s.

W grupie REGULAR każdemu z testerów wyświetlono w kolejności losowej, jednokrotnie, każdą ze 170 sekwencji PVS.

Dla grupy REPEAT, z puli 34 sekwencji SRC wybrano 12 SRC i odpowiadające im sekwencje PVS w każdej grupie jakościowej (A-E). Każdy tester oglądał i oceniał każdą z tych sekwencji PVS trzykrotnie, przy czym sekwencje PVS były prezentowane w losowej kolejności. Łącznie każdy tester obejrzał 180 PVS.

Interfejs do zbierania wyników. Eksperyment został przeprowadzony w laboratorium komputerowym Instytutu Telekomunikacji AGH. Zbieranie danych odbywało się z wykorzystaniem autorskiej platformy testowej, w postaci aplikacji internetowej utworzonej na potrzeby eksperymentu, udostępnionej na serwerze uczelnianym.

Po zalogowaniu się przy użyciu identyfikatora testera, platforma testowa przypisywała automatycznie testera do jednej z dwóch grup eksperymentalnych (REGULAR bądź REPEAT). Bezpośrednio po uruchomieniu eksperymentu, testerowi wyświetlony został pierwszy wybrany w kolejności losowej pojedynczy PVS o czasie trwania 10 s z puli właściwej dla grupy, po czym pojawiał się ekran oceny, zawierający listę jednokrotnego wyboru możliwych ocen według skali ACR. Po dokonaniu oceny wyświetlany był kolejny wylosowany film z puli i proces był powtarzany aż do wyczerpania zbioru PVS. Oceny testerów zostały zapisane w pliku wynikowym na serwerze.

Testerzy. Pula testerów została wyłoniona z populacji studentów AGH zainteresowanych multimediami. W eksperymencie udział wzięło 35 testerów, w tym 7 kobiet oraz 28 mężczyzn, podzielonych na dwie grupy w taki sposób, że grupa REGULAR liczyła 12 osób, zaś grupa REPEAT - 23 testerów. Przypisanie testerów do grup odbyło się na podstawie parzystości lub nieparzystości ostatniej cyfry numeru indeksu, którym tester logował się do platformy testowej, co wpłynęło na brak równoliczności grup poprzez przypadkowy dobór numerów.

4. ANALIZA REZULTATÓW EKSPERYMENTU

Po przeprowadzeniu eksperymentu: (4.1) dobrano narzędzia analizy statystycznej, (4.2) zweryfikowano obserwacje odstające, (4.3) przeanalizowano uzyskane wyniki eksperymentu w odniesieniu do pytania badawczego.

4.1 Dobór narzędzi analizy statystycznej

Przeanalizowano rozkład wyników całej próby badawczej pod kątem jego normalności. Przy założeniach: (1) próby mają charakter losowy, (2) przestrzeń wartości ma zdefiniowany porządek, (3) nieznane są wartości średnie oraz odchylenia standardowe populacji, do weryfikacji rozkładu próby wybrano test Shapiro-Wilka.

Dla próby badawczej uzyskanej ze wszystkich ocen, przy poziomie istotności $\alpha = 0{,}05$ otrzymano w wyniku testu wartość $p = 0{,}00$. Tym samym odrzucono hipotezę H_0, przyjmując hipotezę H_1, mówiącą, że przy zakładanym poziomie istotności wyniku rozkład badanej cechy w populacji nie ma charakteru normalnego. Wobec tego do dalszych analiz zostały wybrane dwa rodzaje testów nieparametrycznych [3]: (1) Test U Manna-Whitheya dla prób niezależnych oraz (2) Test Wilcoxona dla prób zależnych.

4.2 Weryfikacja obserwacji odstających

Weryfikacja obserwacji odstających została przeprowadzona względem poszczególnych testerów, przy użyciu metody opisanej w rekomendacji ITU-T P.913 [5]. Metoda bada korelację średnich wyników testera z odpowiednimi średnimi wynikami wszystkich testerów przy użyciu współczynnika korelacji liniowej Pearsona. Testerzy są odrzucani, jeśli współczynnik odrzucenia spadnie poniżej przyjętego dla testów ACR i sekwencji wideo progu ($r_1 < 0{,}75$).

Dla grupy REGULAR w pierwszym przebiegu procedury dla jednego testera obliczono wartość współczynnika odrzucenia $r_1 = 0{,}71$, co spełnia kryteria jego odrzucenia. Dla pozostałych testerów uzyskano wartości $r_1 > 0{,}75$. W drugim przebiegu procedury, bez testera odrzuconego, wartości r_1 dla pozostałych testerów nie przekroczyły wartości progowej odrzucenia, co zakończyło procedurę.

Dla grupy REPEAT, w pierwszym przebiegu procedury dla dwóch testerów obliczono wartość współczynnika odrzucenia r_1 odpowiednio na poziomie 0,37 oraz 0,74. Dla pozostałych testerów uzyskano wartości $r_1 > 0{,}75$. Odrzucono wyniki jednego testera o najmniejszej wartości r_1, zaś dla pozostałych testerów przeprowadzono drugi przebieg procedury. W tym przypadku dla jednego z testerów uzyskano wynik $r_1 = 0{,}73$, co spełnia kryteria odrzucenia testera. Dla pozostałych testerów uzyskano wartości $r_1 > 0{,}75$. W trzecim przebiegu procedury wartości r_1 dla pozostałych testerów nie przekroczyły progu odrzucenia, co zakończyło procedurę.

Ostatecznie, w wyniku weryfikacji odrzucono oceny otrzymane od jednego testera w grupie REGULAR oraz dwóch testerów w grupie REPEAT. Do dalszej analizy statystycznej dopuszczono oceny zebrane od 11 testerów w grupie REGULAR oraz 21 testerów w grupie REPEAT.

4.3 Analiza wyników eksperymentu

Poniżej przedstawiono wnioski z rezultatów eksperymentu w odniesieniu do pytania badawczego. Wyniki były analizowane w odniesieniu do poszczególnych grup jakości obiektywnej prezentowanych sekwencji. Istniejące opracowania [1][12] koncentrują się głównie na technicznych aspektach sekwencji wideo. W tym przypadku nacisk został położony na wzajemne zależności i

względny układ zmieniających się sekwencji wideo. Uwzględniono również aspekt randomizacji sekwencji opisany w [6].

Wniosek 1. Na poziomie istotności statystycznej α=0,05 dla wszystkich poziomów grup jakościowych wykazano korelację średniej oceny subiektywnej sekwencji (MOS) [7] z ich jakością obiektywną.

Wyniki eksperymentu. W Tab. 1 przedstawiono wartości średnich i wariancji dla poszczególnych grup jakościowych w obydwu próbach badawczych.

Tab. 1. Wartości średnich i wariancji dla grup jakościowych

	Grupa REGULAR (1x)		Grupa REPEAT (3x)	
	MOS	σ	MOS	σ
GJ A	4,44	0,60	4,56	0,42
GJ B	3,71	0,78	3,76	0,68
GJ C	2,84	0,91	2,97	0,56
GJ D	2,05	0,60	2,14	0,58
GJ E	1,29	0,26	1,28	0,29

Weryfikacja statystyczna. Do przeprowadzenia weryfikacji statystycznej otrzymanych wyników wykorzystano nieparametryczny test U Manna-Whitneya, porównując średnie wartości ocen MOS między filmami z sąsiednich grup jakościowych (A vs. B, B vs. C, C vs. D oraz D vs. E). Wyniki testu przedstawiono w tab. 2.

Tab.2. poziom istotności statystycznej p dla testu

Grupa 1	Grupa 2	Poziom istotności *p*
GJ A	GJ B	0,00
GJ B	GJ C	0,00
GJ C	GJ D	0,00
GJ D	GJ E	0,00

Uzyskane wartości p = 0,00 wskazują, że na poziomie istotności α=0,05 hipoteza H_o dla każdej z par grup badawczych w Tab. 2. powinna być odrzucona na rzecz hipotezy H_1, tj. rozkłady ocen subiektywnych MOS w obydwu populacjach są różne. Dla skrajnych grup jakościowych (A v. E) przeprowadzono analizę różnic wariancji MOS za pomocą testu F-Snedecora. W teście uzyskano w obu grupach wartość p=0,00, co wskazuje na poziomie istotności statystycznej 0,05, że wariancja ocen dla grupy o najwyższej jakości jest wyższa niż wariancja ocen dla grupy o najniższej jakości.

Interpretacja wyników. Na zakładanym poziomie istotności statystycznej, w każdej grupie jakościowej oceny subiektywne uzyskane w eksperymencie bezpośrednio korelują się z poziomem jakości obiektywnej wideo - średnia ocena subiektywna MOS jest zawsze wyższa dla grupy jakościowej o wyższej jakości obiektywnej. Ponadto, dla skrajnych grup jakościowych wykazano większe rozproszenie ocen dla sekwencji o wysokiej jakości obiektywnej (GJ A) w porównaniu z filmami o niskiej jakości obiektywnej (GJ E).

Wniosek 2. Na poziomie istotności statystycznej α=0,05 dla najwyższego poziomu jakości obiektywnej wykazano zależność oceny jakości subiektywnej sekwencji od krotności jej wyświetlania.

Wyniki eksperymentu. W grupie badawczej REPEAT przenalizowano wartości ocen MOS dla sekwencji we wszystkich grupach jakościowych, zależnie od krotności wyświetlenia. W Tab. 3 przedstawiono oceny MOS dla pierwszego, drugiego i trzeciego wyświetlenia tego samego filmu.

Tab.3. Średnie wartości MOS w zależności od krotności wyświetlenia danego filmu (PVS) dla grupy REPEAT

	Wyświetlenie 1	Wyświetlenie 2	Wyświetlenie 3
GJ A	4,65	4,57	4,46
GJ B	3,75	3,76	3,78
GJ C	2,99	2,96	2,96
GJ D	2,09	2,15	2,18
GJ E	1,30	1,28	1,26

Weryfikacja statystyczna. Przyjęto procedurę właściwą dla analizy prób zależnych, gdyż dotyczą one kolejnych wyświetleń tych samych sekwencji przez tych samych testerów. Poddano zatem wielokrotnemu pomiarowi tą samą zmienną w określonym odstępie czasu. Bodźcem różnicującym próby jest przyzwyczajenie odbiorcy do powtarzalnej treści. Dla każdej grupy jakościowej przeprowadzono analizę z wykorzystaniem testu kolejności par Wilcoxona dla prób zależnych, porównując różnice między ocenami MOS uzyskanymi w przypadku pierwszego i trzeciego wyświetlenia sekwencji.

Tab. 4 przedstawia uzyskane wartości poziomu istotności statystycznej p dla testu w dwóch grupach jakościowych, dla których uzyskano potencjalnie korelujące się wyniki, tj. dla grupy jakościowej A oraz D.

Tab.4. poziom istotności statystycznej p dla testu w grupie REPEAT – próby zależne

Grupa 1 (krotność 1)	Grupa 2 (krotność 3)	Poziom istotności *p*
Ocena GJ D	Ocena GJ D	0,14
Ocena GJ A	Ocena GJ A	0,00

Dla grupy jakościowej A uzyskana wartość p wskazuje, że na poziomie istotności α=0,05 hipoteza H_0 powinna być odrzucona na rzecz hipotezy H_1, tj. rozkłady ocen subiektywnych MOS w obydwu populacjach są różne. W przypadku grupy jakościowej D na poziomie istotności α=0,05 nie ma podstaw do odrzucenia hipotezy zerowej, natomiast wartość p jest zbliżona do granicznego poziomu istotności 0,05.

Interpretacja wyników. W grupie o najwyższej jakości obiektywnej (A), uzyskano średnią ocenę MOS dla pierwszego wyświetlenia statystycznie istotnie wyższą niż dla trzeciego wyświetlenia. Zatem oceny PVS o wysokiej jakości pogarszają się wraz z kolejnymi wyświetleniami, co może wskazywać na wzrost krytycyzmu testerów wobec treści o obiektywnie wysokiej jakości.

Wniosek 3. na poziomie istotności statystycznej α=0,05 dla najwyższego poziomu jakości sekwencji wykazano zależność oceny subiektywnej od jakości poprzednio wyświetlonej sekwencji.

Wyniki eksperymentu. W obydwu grupach badawczych porównano wartości MOS dla wszystkich grup jakościowych w zależności od grupy jakościowej filmu poprzedzającego. Otrzymane wyniki przedstawiono w Tab. 5 oraz Tab. 6.

Tab.5. Średnie wartości MOS w zależności od jakości obiektywnej sekwencji poprzedzającej dla grupy REGULAR

	GJ sekwencji poprzedzającej				
GJ filmu bieżącego	A	B	C	D	E
GRUPA A	4,48	4,41	4,39	4,58	4,35
GRUPA B	3,79	3,78	3,79	3,45	3,72
GRUPA C	2,88	2,99	2,74	2,87	2,72
GRUPA D	2,13	2,13	1,95	1,96	2,05
GRUPA E	1,27	1,30	1,29	1,37	1,24

Tab.6. Średnie wartości MOS w zależności od jakości obiektywnej sekwencji poprzedzającej dla grupy REPEAT

	GJ sekwencji poprzedzającej				
GJ filmu bieżącego	A	B	C	D	E
GRUPA A	4,53	4,47	4,54	4,42	4,31
GRUPA B	3,74	3,70	3,61	3,67	3,69
GRUPA C	3,03	2,98	2,93	2,95	2,92
GRUPA D	2,34	2,19	2,14	2,04	2,07
GRUPA E	1,44	1,31	1,36	1,28	1,34

Weryfikacja statystyczna. Dla każdej grupy jakościowej przeprowadzono analizę testem U Manna-Whitneya, porównując różnice między ocenami MOS uzyskanymi w przypadku, gdy poprzedniczkami były sekwencje o skrajnej jakości obiektywnej (bardzo wysokiej z grupy jakościowej A oraz bardzo niskiej z grupy jakościowej E). Test statystyczny wykonano osobno w obydwu grupach REGULAR oraz REPEAT. Tab.7 przedstawia uzyskane wartości poziomu istotności statystycznej p dla testu w grupie REPEAT.

Tab.7. poziom istotności statystycznej p dla testu w grupie REPEAT

Grupa 1 (poprzednik GJ A)	Grupa 2 (poprzednik GJ E)	Poziom istotności p
GJ A	GJ A	0,04
GJ D	GJ D	0,00

W grupach jakościowych A oraz D uzyskana wartość p wskazuje, że na poziomie istotności α=0,05 hipoteza H_0 dla każdej z par grup badawczych powinna być odrzucona na rzecz hipotezy H_1 - rozkłady ocen subiektywnych MOS w obydwu populacjach są różne. W pozostałych przypadkach na poziomie istotności α=0,05 nie ma podstaw do odrzucenia hipotezy zerowej o równości rozkładów.

W grupie REGULAR, dla wszystkich par uzyskana wartość p (≥0,42) wskazuje, że na poziomie istotności α=0,05 nie ma podstaw do odrzucenia hipotezy zerowej o równości rozkładów.

Interpretacja wyników. Odnosząc uzyskane wyniki analizy statystycznej do otrzymanych wyników eksperymentu, należy zauważyć, że wykazana prawidłowość lepszej średniej oceny subiektywnej MOS w przypadku, gdy film poprzedzający jest zdecydowanie lepszej jakości została uzyskana na każdym poziomie jakości obiektywnej sekwencji. Co prawda istotność statystyczną uzyskano jedynie w grupie REPEAT dla grup jakościowych B oraz E, natomiast uzyskane powtarzające się lepsze wyniki w grupie z poprzednikiem bardzo wysokiej jakości wskazują na celowość dalszych badań w tym zakresie.

5. WNIOSKI I DALSZE REKOMENDACJE

Uzyskane wyniki wskazują, że zidentyfikowane zależności istotne statystycznie dotyczą przede wszystkim sekwencji o najwyższej jakości. Dalsze kierunki badawcze mogą dotyczyć takich aspektów jak:

- badania wpływu wzorców odtwarzania w odniesieniu do obecnego stanu wiedzy [1],
- pogłębienie analizy względem sekwencji z bardziej zdegradowanych grup jakościowych,
- analiza korelacji wzajemnych między czynnikami (np. czynnik ludzki vs. struktura prezentacji filmów),
- Korelacja z innymi czynnikami (np. środowiskowymi, społecznymi, ludzkimi),
- pogłębienie analizy poprzedzających sekwencji (np. analiza n poprzedników (n>1)).

LITERATURA

[1] Bampis, C. G., Li, Z., Moorthy, A. K., Katsavounidis, I., Aaron, A., & Bovik, A. C. 2017. Study of temporal effects on subjective video quality of experience. *IEEE Transactions on Image Processing*, *26*(11), 5217-5231.

[2] Brunnström, et al. 2013. Qualinet white paper on definitions of quality of experience.

[3] Corder, G. W., & Foreman, D. I. 2014. *Nonparametric statistics: A step-by-step approach*. John Wiley & Sons.

[4] ITU-T. Recommendation ITU-T P.800: Methods for subjective determination of transmission quality, International Telecommunication Union, 1996

[5] ITU-T. Recommendation ITU-T P.913: Methods for the subjective assessment of video quality, audio quality and audio-visual quality of Internet video and distribution quality television in any environment, International Telecommunication Union, 2016.

[6] Juszka D. and Papir Z. 2015. "A study on order effect in a subjective experiment on stereoscopic video quality," *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, Pilos, Greece, 2015, pp. 1-6, doi: 10.1109/QoMEX.2015.7148087.

[7] Kilkki, K. 2008. Quality of experience in communications ecosystem. J. Univers. Comput. Sci., 14(5), 615-624.

[8] Raake, A., Egger, S. 2014. Quality and Quality of Experience. In: Möller, S., Raake, A. (eds) Quality of Experience. T-Labs Series in Telecommunication Services. Springer, Cham.

[9] Rassool R. 2017. "VMAF reproducibility: Validating a perceptual practical video quality metric," IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Cagliari, pp. 1-2.

[10] Reiter, U. *et al.* 2014. Factors Influencing Quality of Experience. In: Möller, S., Raake, A. (eds) Quality of Experience. T-Labs Series in Telecommunication Services. Springer, Cham.

[11] Schatz, R., Hoßfeld, T., Janowski, L., & Egger, S. 2013. From packets to people: quality of experience as a new measurement challenge. *Data traffic monitoring and analysis: From measurement, classification, and anomaly detection to quality of experience*..

[12] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, *13*(4), 600-612.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

UNSUPERVISED CATEGORY LEARNING IN PERCEIVED VIDEO QUALITY

Andres Vejar[1]; Tomasz Konazyński[2]; Dawid Juszka[3]; Mikołaj Leszczuk[4]

Institute of Telecommunications, AGH University of Krakow, Poland.
[1]avejar@agh.edu.pl, [2]konaszyn@agh.edu.pl, [3]dawid.juszka@agh.edu.pl, [4]mikolaj.leszczuk@agh.edu.pl

DOI: 10.15199/59.2023.4.74

Abstract: **In video quality evaluation, the perceived quality is ranked by the participants using a categorical scale of five levels. To study the category learning dependency, the participants were divided into learners and no-learners, with respect to their classification accuracy. An analysis of the performance of the human unsupervised learning from machine learning models is presented in order to study the effects of category learning in the video assessment.**

Keywords: **Category learning, Perceived video quality, Machine Learning Classification**

1. INTRODUCTION

Classification is the process of assigning a sample to a specific category. For machine learning classification, the concept of unsupervised learning consists on the construction of algorithms that can learn to classify samples without providing a "train set" of samples and their corresponding category labels. We can consider the process of video evaluation as a classification task, where the perceived quality is rated by the participants using categories, generally labeled "excellent", "good", "fair", "poor" and "bad". Multiple videos are presented to the user, in random order with respect to the quality of the sample. From this perspective, the traditional process of video quality evaluation conforms with the problem of unsupervised category learning.

The process of selecting a specific category is not trivial, specially when the user needs to score on the category limits, because one video considered 'bad' at the beginning of the video evaluation, can be underrated or overrated, after understanding the quality diversity of the whole video set. This effect can variate from evaluator to evaluator, but it is unavoidable, given that learning is an innate ability [1], generally impossible to disentangle from the activities performed. The "passive" category learning task used in psychological studies [2] aligns precisely with the description of the classical video quality evaluation experiment [3].

From the cognitive perspective, there is an important effect produced by category learning, and it is called categorical perception [4, 5], where a within-category compression effect is observed after category learning. For video quality evaluation, if the categorical perception effect is induced over the trial, then the samples belonging to the same categories will be perceived near. In [6], an experiment is presented where behavioral and electrophysiological data were used to analyze the changes in perception. In the study, perceptual effects were observed positively for learners (individuals that categorize with 80% accuracy). To explain these effects, neural network models of induced categorical perception can be constructed [7, 8]. This effect is not limited to living agents, the category learning is also observed for pure machine learning models [9].

The motivation of this work is to associate the process of video quality evaluation with unsupervised classification. This association involves machine learning and the general concept of category learning. To quantify the effects of learning from video quality, we analyze the user data from two types of video experiments. For the analysis, the participants of the experiments can be divided into learners and no-learners, with respect to their classification accuracy.

2. MATERIALS AND METHODS

Traditionally, a link between the subjective concept of video quality, and the subjective rating of the users is studied. The mean opinion score (MOS) for an experiment is associated to objective video quality indicators (VQI) that consist on quantitative, analytic calculations from the image sequences. Let the variable $x \in \mathcal{X}$ represent a video sequence. Generally $\mathcal{X} = [\![0\,..\,255]\!]^N$, where $N = T \times W \times H \times C$ and $[\![0\,..\,255]\!]$ corresponds to the integer interval of intensities for each 8-bit pixel. The video geometry is $W \times H$ pixels, with C color channels and T frames. To construct experiment samples, a source video (SRC) is processed to obtain different quality levels, using hypothetical reference circuits (HRC), generally corresponding to a degradation function $f : (\Theta \times \mathcal{X}) \to \mathcal{X}$, with parameters $\theta \in \Theta$. The resulting samples are called processed video sequences (PVS). For a SRC x_0, and a parameter vector θ_1, we obtain the PVS $x_1 = f(\theta_1, x_0) = x_0 + \varepsilon$. Basic quality indicators can be derived with L1 or L2 norms, for example the mean squared error: $\mathrm{MSE}(x_0, x_1) = (1/N)||x_0 - x_1||_2^2 = (1/N)||\varepsilon||_2^2$.

VQI can be classified in three types:

1. The **full reference** quality indicator $g : (\mathcal{X} \times \mathcal{X}) \to \mathbb{R}$, that requires the SRC and the PVS.
2. The **reduced reference** quality indicator $\breve{g} : (\mathbb{R}^H \times \mathcal{X}) \to \mathbb{R}$, requires only a projection of the SRC and a full PVS. The projection can be defined as $h : \mathcal{X} \to \mathbb{R}^H$, where the dimension H is lower

than the dimension of original input space $\mathcal{X}$.

3. The **no reference** quality indicator $\bar{g} : \mathcal{X} \to \mathbb{R}$, it uses only the PVS as input.

Full reference allows to analyze deeply the differences between SRC and PVS, with considerable use of computational resources. Reduced reference allows to decrease the use of computational resources, still providing information about the SRC. No reference VQIs can require a small use of computer resources but they provide very limited reliability.

To asses the result of one video quality experiment with U users and M PVS, the MOS is calculated as $\text{MOS}(x_j) = (1/U)\sum_{u=1}^{U} r_{uj}$, for each PVS x_j, $j = 1 \dots M$. The rating of the user u with respect to the PVS index j is $r_{uj} \in [\![1 \,..\, 5]\!]$, with categorical labels: "excellent"$\mapsto$ 5, "good"$\mapsto$ 4, "fair"$\mapsto$ 3, "poor"$\mapsto$ 2 and "bad"$\mapsto$ 1.

To create accurate video quality models, a set of calculated VQIs can be used as features of a predictive video quality assessment model:

$$\widehat{\text{MOS}(x_j)} = Q(v_1, v_2, \dots, v_Q), \tag{1}$$

with $v_q = g_q(x_{0j}, x_j)$ is the calculated value of the VQI g_q, $q = 1 \dots Q$, for the SRC x_{0j} and the PVS x_j.

The predictive model (1) allows to construct PVS to test different hypothesis in video quality evaluation. Its main applications are: **stream selection** to send the best quality video possible, **quality monitoring** to track the estimated quality of the streams, **video restoration** to improve the quality of the video, and **encoder optimization**, to encode to the best possible quality [10].

For a given experiment, and a collection of PVS $X = (x_1, x_2, \dots, x_M)$, each participant u of the experiment classify sequentially a random permutation $\boldsymbol{\pi}_u$ of the indexes of X. Individual results are collected in the matrix $Y_u \in \{0,1\}^{M\times 5}$, where $Y_u[i,j] = 1$ if the user u classify the PVS i in the category j.

If the original quality levels for X are provided in the matrix $R \in \{0,1\}^{M\times 5}$ then the accuracy of unsupervised classification for the user u is:

$$A_u = \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{5} R[i,j]\cdot Y_u[i,j] \tag{2}$$

The experiment users can be separated in learners and no-learners for some level of accuracy $\alpha \in (0,1)$. Other measures can be used to quantify the quality of the of the unsupervised classification, for example, the confusion matrix or the cross-entropy.

3. RESULTS AND DISCUSSION

We consider, as data sources, two types of experiments designed and executed by the members of the Video Quality of Experience group of the Institute of Telecommunications, AGH University of Krakow:

1. Experiment type 1x: considers 34 unique SRC, and 5 quality levels. Total: 170 unique PVS evaluated per user.

2. Experiment type 3x: only 12 unique SRC are selected for the 34 unique SRC of Experiment type 1x, with 5 quality levels. In total, 60 unique PVS. The user assess every video three times, in a random sequence. Total: 180 PVS evaluated per user.

The experiment was conducted with a carefully curated video dataset, using the state-of-the-art, full reference VMAF algorithm to construct five levels of quality from the SRCs. In this work we consider the degradation function as a black-box, therefore we do not use prior information about the process of creation of each PVS.

The videos variate in geometry and duration. In order to account for the geometry variations and any slight change in video duration or number of frames with respect to the SRC, we calculate a scaled bit per pixel s as follows:

$$s = \frac{1}{T}\frac{rt}{WH} \tag{3}$$

where T is the number of frames of the PVS, r (bit/s) is the bitrate, the number of pixels is $W \times H$, and the total video duration is t (s).

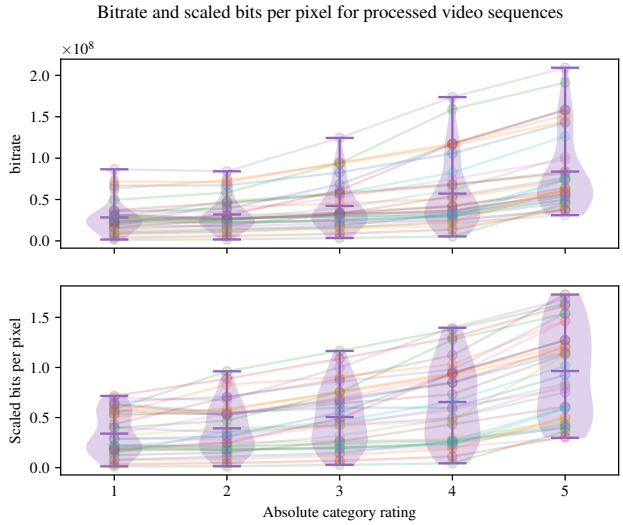


Figure 1: Bitrate and scaled bits per pixel on processed video sequences (PVS). Light lines represents the 34 unique SRC, with theirs PVS lying on different quality levels (1:lowest, 5:highest).

In Figure 1 are presented two plots that display the bitrate with respect to the absolute category rating scale for the 170 PVS considered for the experiments. The bottom plot presents the scaled bits per pixel, where it is possible to appreciate an increase in bits per pixel with respect to the aggregate values of the whole set. This aggregate view presents high variance within classes. The top plot of Figure 1, shows the bitrate, and it reflex better the differentiation between categories, with a small variance than for the bottom plot. The lines between the categories, for both plots, connects all the quality version PVS for a specific SRC. We can appreciate that for the lower classes, some

of the aggregate bit values don't follow the order imposed by the dataset.

The experiments collect the duration of the assessment for each participant and for each sample. This information is important to quantify variations in the response time of the user with respect to the classification task. Exist evidence in the literature about modeling response times as a lognormal distribution [11, 12, 13, 14]. In [15], a multiagent system formulated using statistical mechanics is constructed to explain the lognormality in human activities, e.g. in call-centers service times [16].

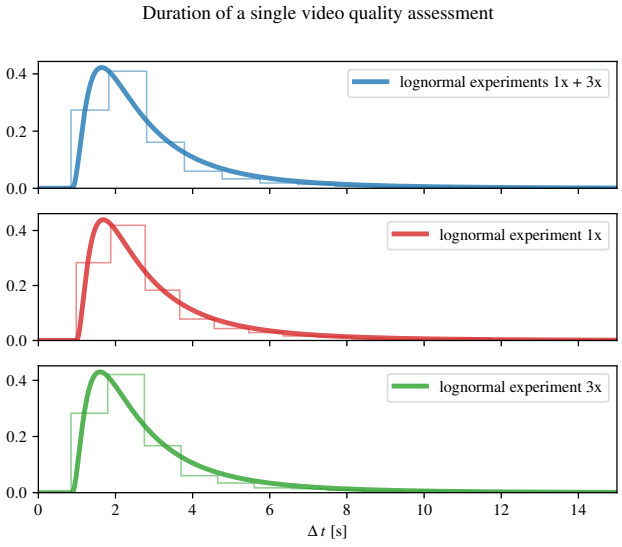


Figure 2: Histogram and lognormal PDF estimation of the duration of the assessment of a processed video sequence. Top row: global duration for experiment type 1x and 3x. Middle row: duration for experiment type 1x. Bottom row: duration for experiment 3x.

The duration of the classification of one PVS by the user is presented in Figure 2. The data are likely to follow a lognormal distribution. The top row, in blue, collect the data for both experiments in a histogram, the continuous thick line represents the estimated probability density function (PDF) from a lognormal model:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right) \qquad (4)$$

The middle row of Figure 2 present, in red, the histogram and the PDF estimation for the experiment of type 1x. Similarly, the bottom row, in green, present the histogram and the PDF estimation for the experiment of type 3x. Comparing the PDF for exp. type 1x and 3x, we can appreciate that the mean and the variance is higher in the 1x case than in the 3x case. This variation can be produced given the higher diversity of cases in exp. type 1x, and also by the repetition of samples in exp. type 3x.

To test the effects of learning over the response time, i.e. the duration of the video assessment task, the user duration data was separated in five periods. In Figure 3, the duration distributions for the two experiment types are presented. In both cases, the overall duration is reduced after the first two periods (after 40% of the experiment). For experiment type 3x, where the set of unique PVS is repeated three times, the variance is minimized in the last period, probably a consequence of the learning process and the memory effect caused by repetition.

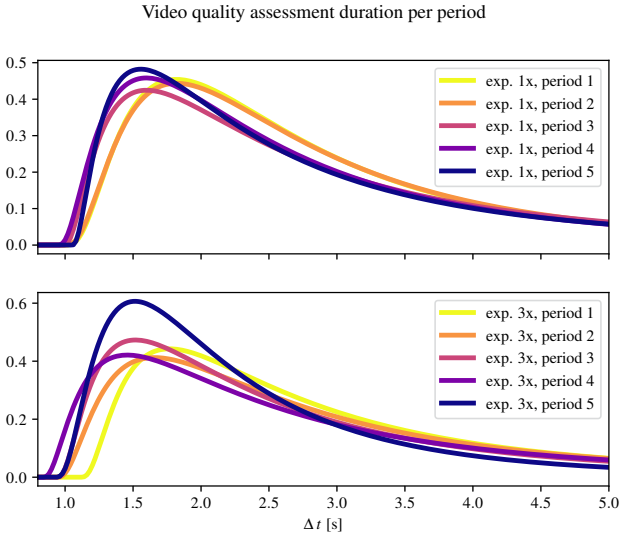


Figure 3: Duration of the assessment of a processed video sequence for five chronological periods. Top row: experiment type 1x. Bottom row: experiment type 3x.

To calculate the error in the video quality category learning, we consider both positive (under estimation) and negative errors (over estimation). To visualize the variations on the user errors for different chronological periods, a kernel density estimation with a gaussian kernel is performed. For a dataset $\{x_i\}_{i=1}^{N}$, the kernel density estimation of the PDF $f(x)$ is defined by:

$$\hat{f}_w(x) = \frac{1}{Nw} \sum_{i=1}^{N} K\left(\frac{x - x_i}{w}\right) \qquad (5)$$

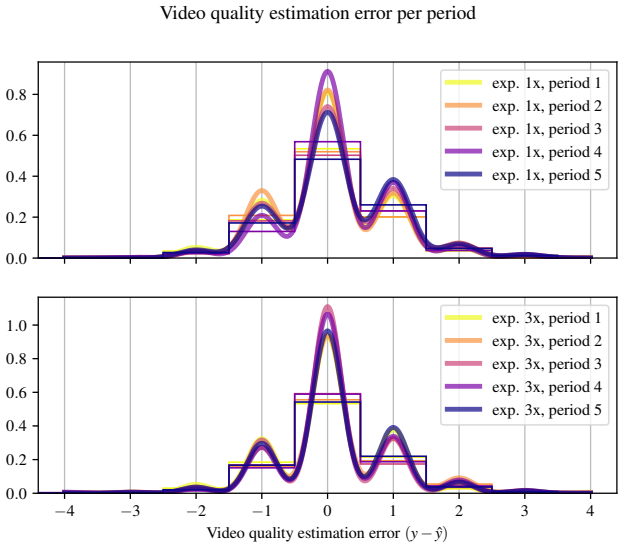


Figure 4: Video quality assessment error of a processed video sequence for five chronological periods. Top row: experiment type 1x. Bottom row: experiment type 3x.

In (5), $w > 0$ is the bandwidth parameter, and K is the kernel function, in this case the standard normal distribution. The results are presented in Figure 4. In general, it is noticeable how under-estimation error (positive) is more probable than the over-estimation error, independent of the type of experiment. With respect to the learning evolution, the probability of zero error increases after half of the experiment (period 4), but it is reduced in a high extend in both cases for the last period (last 20% of the experiment). This reduction can be produced by the fatigue of the users. It is possible to appreciate that the variations in the error over periods are smaller in the experiment type 3x. This stability can be generated by the repetition of the PVS and by the reduction of PVS diversity in exp. type 3x.

4. CONCLUSIONS

In this work we analyze the process of video quality evaluation, and we compare it with respect to the unsupervised classification task in machine learning. Considering the process of learning from cognitive science, it is natural that the participants of the video task will modify their behavior along with the trial. With respect to response times, a reduction is present in the two types of experiments analyzed. With respect to the performance of the users to classify correctly, it is observed that there is a general bias to under-estimate the true quality of the PVS. The most accurate period occurs during the middle part of the experiment. During the last 20% of the experiment the accuracy falls, effect probably induced by fatigue.

Further work can consider the modeling of the video quality evaluation in more detail, with respect to the order of presentation of PVS for each participant of the experiment. Also, a comparison of the quality of the datasets with another algorithms like multi-scale structural similarity can be of great interest to verify the results provided by the VMAF algorithm.

An interesting extension is to provide the video quality assessment experiment as an active-learning task: A navigable selection window with small clips or snapshots of the videos to evaluate is presented to the user. The user then decides which sample to evaluate each time. This is motivated by the empirical evidence of active learners achieving better classification accuracy than passive learners.

REFERENCES

[1] James L. Gould and Peter Marler. Learning by instinct. *Scientific American*, 256(1):74–85, 1987.

[2] Todd M. Gureckis and Douglas B. Markant. Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, 7(5):464–481, 2012. PMID: 26168504.

[3] ITU-T. Recommendation ITU-T P.910: Subjective video quality assessment methods for multimedia applications. Recommendation, International Telecommunication Union, 2022.

[4] Kenneth R Livingston, Janet K Andrews, and Stevan Harnad. Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(3):732, 1998.

[5] Robert L Goldstone, Yvonne Lippa, and Richard M Shiffrin. Altering object representations through category learning. *Cognition*, 78(1):27–43, 2001.

[6] Fernanda Pérez-Gay Juárez, Tomy Sicotte, Christian Thériault, and Stevan Harnad. Category learning can alter perception and its neural correlates. *PLOS ONE*, 14(12):1–29, 12 2019.

[7] RI Damper and SR Harnad. Neural network models of categorical perception. *Perception & psychophysics*, 62(4):843–867, 2000.

[8] Christian Thériault, Fernanda Pérez-Gay, Dan Rivas, and Stevan Harnad. Learning-induced categorical perception in a neural network model, 2018.

[9] Laurent Bonnasse-Gahot and Jean-Pierre Nadal. Categorical Perception: A Groundwork for Deep Learning. *Neural Computation*, 34(2):437–475, 01 2022.

[10] P. Topiwala, W. Dai, J. Pian, K. Biondi, and A. Krovvidi. VMAF and variants: towards a unified VQA. In Andrew G. Tescher and Touradj Ebrahimi, editors, *Applications of Digital Image Processing XLIV*, volume 11842, page 118420F. International Society for Optics and Photonics, SPIE, 2021.

[11] Wim J. van der Linden. A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31(2):181–204, 2006.

[12] Jean-Paul Fox, Rinke Klein Entink, and Wilm van der Linden. Modeling of responses and response times with the package cirt. *Journal of Statistical Software*, 20(7):1–14, 2007.

[13] Sandip Sinharay and Peter W. van Rijn. Assessing fit of the lognormal model for response times. *Journal of Educational and Behavioral Statistics*, 45(5):534–568, 2020.

[14] Paul De Boeck and Minjeong Jeon. An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, 10, 2019.

[15] Stefano Gualandi and Giuseppe Toscani. Human behavior and lognormal distribution. a kinetic description. *Mathematical Models and Methods in Applied Sciences*, 29(04):717–753, 2019.

[16] Stefano Gualandi and Giuseppe Toscani. Call center service times are lognormal: A Fokker–Planck description. *Mathematical Models and Methods in Applied Sciences*, 28(08):1513–1527, 2018.

SUBIEKTYWNA OCENA JAKOŚCI ALGORYTMÓW PODSUMOWANIA WIDEO: PODEJŚCIE OPARTE NA CROWDSOURCINGU

SUBJECTIVE QUALITY ASSESSMENT OF VIDEO SUMMARISATION ALGORITHMS: A CROWDSOURCING APPROACH

Avrajyoti Dutta[1]; Mikołaj Leszczuk[2]; Dawid Juszka[3]; Michał Grega[3]

[1]AGH University of Krakow, avrajyoti.dutta@agh.edu.pl
[2]AGH University of Krakow, mikolaj.leszczuk@agh.edu.pl
[3]AGH University of Krakow, dawid.juszka@agh.edu.pl, michal.grega@agh.edu.pl

DOI: 10.15199/59.2023.4.75

Streszczenie: Streszczanie wideo to sposób selekcji najbardziej istotnych i informatywnych fragmentów nagrań wideo. Jego celem jest utworzenie takiej wersji nagrania, która będzie skondensowana i pozwoli szybko zapoznać się z najważniejszymi treściami. W artykule prezentujemy eksperyment subiektywny przeprowadzony metodą crowdsourcingową. Zadaniem uczestników było ocenienie kilku algorytmów streszczania wideo. Wyniki umożliwią ulepszenie algorytmów. Wskazaliśmy także praktyczne ograniczenia stosowania crowdsourcingu.

Abstract: Video summarization is a technique that extracts the most relevant and informative segments of a video to create a condensed version that is faster to watch. In this paper, we present a crowdsourcing subjective experiment in which summarizations of video clips were assessed. The aim of the experiment was to determine the accuracy of the video summarization algorithm. The results of the evaluation will be used to improve the video summarization algorithm. The experiment also revealed some limitations of the crowdsourcing method.

Słowa kluczowe: crowdsourcing, eksperyment subiektywny, ocena jakości wideo, streszczenie wideo, treść generowana przez użytkownika.

Keywords: Crowdsourcing, Subjective Experiment, Video Quality Assessment, Video Summarization, User-generated content.

1. INTRODUCTION

The advent of the digital age has resulted in an explosion of video content, with millions of hours of videos being uploaded to the internet every day. This vast amount of information has made it increasingly difficult for users to find and consume relevant content. Because the perspective of the end user directly affects the Quality of Experience (QoE) of multimedia applications, determining the level of user satisfaction or aggravation requires subjective user research [1]. Evaluating the efficacy of video summarization algorithms can prove to be a complex undertaking, given its subjective nature and dependence on diverse factors, including but not limited to the context, purpose, and intended viewership. Video summarization is the process of creating a summary of a video clip by extracting the most important parts from it. It is a difficult task that requires a high degree of expertise and time. Recently, crowdsourcing has been proposed as a possible solution to this problem. The idea is to employ many people to work on the task in order to generate summaries in a short amount of time.

Video summarization has become an essential component of modern multimedia technology. The process of summarizing short videos can be a challenging and time-consuming task. As a result, there is a growing need for summarization of video content. Video summarization is a complex task and has been approached in many ways in the past. In recent years, various summarization methods have been proposed for summarization of videos. The importance of video summarization is increasing due to the growing amount of video content available on the internet. Crowdsourcing has been shown to be a useful tool for video summarization as it allows for the aggregation of diverse opinions and perspectives.

In this research, we present a crowdsourcing experiment for scenario-wise video summarization. The experiment aimed to evaluate the effectiveness of crowdsourcing for video summarization and to investigate the impact of different scenarios on the quality of the summaries. This paper presents an experiment to test the feasibility of this technique. We designed an experiment where we hired individuals to watch and summarize video clips. The experiment was conducted over a period of twelve weeks. This was conducted to evaluate the effectiveness of the crowdsourcing technique in generating video summaries. We created a dataset (i.e., sample of videos) of video clips, and participants were asked to watch the video clips and generate a voting poll of summaries. We compared the results of our approach to other video summarization techniques, such as scenario-wise algorithms.

Our results show that the crowdsourcing technique is effective in summarizing video clips, and it outperforms other methods. We evaluated the quality of the summaries using a set of metrics. The results showed that the crowdsourcing technique is effective in generating high-quality summaries. The results of the experiment showed that the crowdsourcing technique is more effective in terms of accuracy and time than traditional summarization methods.

Therefore, on the creation of a new indicator for the detection of user-generated content or UGC, as well as the general extension of this experiment's solution. The effectiveness of the abstract system itself will be evaluated.

2. TEST BACKGROUND

2.1 CROWDSOURCING

Crowdsourcing is a technique where a task is divided into small parts and distributed among a large group of individuals. The crowd can be utilized to perform tasks that are difficult or impossible for machines to do.

This involves outsourcing tasks to a large group of people, typically via an online platform [2]. This approach involves disseminating a summary to a broad audience via an internet-based platform, followed by the collection of feedback from said audience. Although this approach has the potential to yield a significant volume of information within a brief timeframe, the caliber of the feedback may fluctuate. The individuals complete the task independently, and the results are combined to achieve the outcome. Crowdsourcing has been used successfully in various applications, including image labeling, data annotation, and speech recognition. Crowdsourcing can also have broader implications for content creation and curation. Human-centered experiments on perceptual quality are thought to be the most reliable way to assess the quality of experience [3]. Crowdsourcing can be used to create user-generated content, which can be a powerful tool for engagement and community building. Furthermore, crowdsourcing can also be used to curate content for specific audiences, such as educational videos for children or news highlights for busy professionals. Crowdsourcing has emerged as a viable alternative to traditional summarization techniques. This approach has several advantages, including the ability to obtain summaries from a diverse group of people, at a relatively low cost, and within a short period.

In our approach, we hired individuals to watch and summarize video clips. The individuals were provided with a set of guidelines and asked to summarize the video clips based on the guidelines. In recent years, crowdsourcing has gained popularity as an effective way to generate summaries from video clips.

However, there are several challenges associated with crowdsourcing for video summarization. One of the challenges is ensuring the quality of the summaries produced by the crowd. Another challenge is the potential for bias and manipulation by the crowd, which can result in low-quality summaries.

Despite these challenges, the potential benefits of crowdsourcing for video summarization are significant. Crowdsourcing can provide a more comprehensive and diverse perspective on video content, leading to summaries that better capture the needs and preferences of different users. Moreover, crowdsourcing can also reduce the time and cost associated with video summarization, making it a more efficient and accessible process.

2.2 ALGORITHM OF SUMMARIZING VIDEO CLIPS

In total we have developed 5 different algorithms written in Python for summarization of the video clips. Each algorithm accepts the input video file and the desired length as parameters and returns the summary.

For each of the following algorithms the video is divided into scenes using the pySceneDetect package [4]. For each scene an average Spatial Activity (SA) and Temporal Activity (TA) are calculated using the agh-vqis package [5,6]. A multiplication of these parameters was calculated for each scene, to depict both the SA and the TA of a single scene with a single value (hereafter referred to as a "Coefficient"). Moreover, for each scene a detector of User Generated Content (UGC) was used to automatically classify, if a given scene is User- or Professionally generated. Further details on this UGC detector are provided in [7].

The first algorithm ("A") was a baseline algorithm, described in detail in [8]. It is based on TA and SA only. The scenes in the video were ordered in a descending order by the Coefficient, thus promoting the scenes of the higher activity overall. From this list scenes were selected from the top to be included into the summary until the desired summary length was achieved. Finally, the selected shots were sorted in the chronological order of appearance in the original video.

The second algorithm ("B") utilized the UGC detector in order to promote the UGC scenes in the summary. Then, the remaining scenes were divided into two categories - the UGC scenes and the non-UGC (professional) scenes. Each was sorted in the descending order by the Coefficient. First, the scenes from the top of the UGC list were selected until the desired length of the summary was reached. If there was insufficient number of the UGC scenes - scenes from the top of the non-UGC list were selected until the desired length of the summary was reached. As before, the selected shots were sorted in the chronological order of appearance in the original video. We tested the hypothesis that the UGC content conveys more important information than non-UGC content.

The third algorithm ("C") also utilized the UGC detector. First, the initial shot of the input video was included into the summary - as usually it conveys important information on the content of the video. In this strategy the shots were divided into UGC and non-UGC lists and sorted by the coefficient as before. However, the shots were selected in an alternating way from the UGC and non-UGC lists until the desired duration was reached. Finally, the selected shots were sorted in the chronological order. Here, we tested the hypothesis that the summary should contain a mix of UGC and non-UGC content.

The fourth algorithm ("D") was identical to the second one ("B") with the only one difference that the first shot of the input video was always included into the summary. We have included this scenario, as the results of the "C" algorithm were promising, possibly thanks to inclusion of the first shot.

The fifth algorithm ("F") was similar to the fourth ("D") and second ("B") scenarios. First shot was always included. Afterwards, this algorithm promoted non-UGC shots in the analogous way that the "B" scenario promoted the UGC content. Here, we have tested the hypothesis that the non-UGC shots convey more important information that the UGC ones (this is the opposite of the hypothesis tested in "D").

For each algorithm, after the scenes were selected, the output video was automatically edited in order to concatenate the selected clips (in their chronological order) and to include the proper audio track. This was performed using the popular ffmpeg program.

The creation and improvement of visual indicators and summarizing algorithms will be the focus of industrial research. The application's method for summarizing video sequences is based on a variety of visual indicators. The increased benefit of employing the UGC indicator will be measured in this way.

3. METHODOLOGY

The experiment was conducted with a group of more than 45 participants. The participants were asked to watch 30- and 45-second video clips and to assess summaries of the contents in a voting poll. To evaluate the effectiveness of the proposed technique, we conducted an experiment using a dataset of 24 videos. The video clips were chosen based on their length and content. We compared our technique with presentation order and random selection methods. The videos were segmented into smaller clips using a content-based segmentation method. The videos were also selected to represent different levels of complexity and visual richness. We also conducted a subjective evaluation by asking human evaluators to summarize the summaries based on their usefulness, coherence, and completeness. The proposed experiment involves the following steps:

Step 1: Selection of Video Clips - We select a set of video clips from various domains, such as news, sports, entertainment, etc. The video is segmented into smaller clips based on different criteria such as content-based segmentation or user-defined segmentation. The video clips should be short (30 and 45 seconds) and contain a clear narrative structure. We use algorithms for the video clips. For each video, two different scenarios were defined, and participants were randomly assigned to one of the scenarios. The experiment has a total of five scenarios.

Step 2: Crowdsourcing Platform - We use an online crowdsourcing platform, such as our own website (excluding platforms like Amazon Mechanical Turk, MicroWorkers etc), to recruit participants for the experiment. The participants are required to have a good command of the Polish language and access to a stable internet connection with their laptop or tablet.

Step 3: Task Description - We provide the participants with a task description that explains the purpose of the experiment and provides clear instructions on how to provide a summary of the video clip by giving a voting poll.

Step 4: Video Summarization - The participants are presented with a short video clip and are asked to watch the entire video and provide a summary of the video in their vote. The selected votes are evaluated by human annotators through a crowdsourcing platform.

Step 5: Data Analysis - We collect the summaries provided by the participants and analyze them using pie charts and bar graphs. After that, we assess the summaries and create a condensed summary of the video clip.

Step 6: Evaluation - We evaluate the quality of the summaries generated by the participants and compare them to the summaries generated by state-of-the-art summarization algorithms. We collected the summarized voting data and analyzed them to determine the effectiveness of the crowdsourcing technique. We use metrics such as statistical formulas to evaluate the statistical analysis.

4. RESULTS

The results showed that the participants were able to accurately assess summarizations of the video clip. We compared the results of our approach with other video summarization techniques. We evaluated the summaries based on their relevance and completeness. The statistical analysis of subjective opinion scores was conducted to ascertain the summary's overall status. The statistical significance level of the proposed technique was 53.4%, indicating that the summaries were of reasonable quality. The second video received more votes than the first one; specifically, 58% of votes were for the second video that was presented. Analyzing the output produced by these algorithms is one technique to make this improvement. Through the analysis of algorithmic output, it is possible to identify perceived patterns and trends that may serve as indicators of potential areas for algorithmic enhancement. The aforementioned data can subsequently be utilized to adjust the algorithm's parameters or to devise novel methodologies for data processing, with the aim of enhancing its precision.

Apart from identifying potential areas of enhancement, the examination of algorithmic outcomes can also facilitate our comprehension of the algorithms' functionality and the variables that impact their performance. The subjective evaluation results also showed that the summaries generated by the proposed technique were rated as more useful, coherent, and complete than the summaries generated by the other methods. This data has the potential to facilitate the development of novel algorithms with enhanced efficacy, or to optimize the production processes of extant algorithms. The results of the experiment showed that crowdsourcing can be an effective tool for assessment of video summarization.

In general, the evaluation of algorithmic outcomes is a potent instrument for enhancing the precision and efficacy of algorithms. Through the analysis of the algorithms' output, significant insights can be obtained regarding their functionality and the variables that impact their behavior. With the aforementioned data at our disposal, we can proceed to devise novel algorithms or enhance preexisting ones, with the aim of generating more precise and effective resolutions to complex challenges. By offering a thorough analysis of the major problems encountered in QoE crowdtesting and the related solutions on the example of QoE evaluation for video, we offered in this contribution best practices for QoE crowdtesting. For effective QoE crowdtesting campaigns, we specifically addressed the design, execution, and reliability evaluation. We have put up a list of recommended practices for crowdsourcing QoE testing [1]. The importance of user satisfaction with internet applications and services is rising, and as a result, service providers are giving QoE more and more thought [9]. The strategy of content-based segmentation involves the utilization of specific criteria and analysis of content characteristics in order to divide videos into individual clips that exhibit diverse levels of complexity and visual richness. The method aims to enhance the manageability of videos for subsequent processing or analysis.

5. DISCUSSION

We evaluated the 'A' scenario in relation to five distinct scenarios: B, C, D, F, and D-45. Our findings indicate that the "D" scenario received 80 votes in favor of the second video. On the other hand, the "F" scenario yielded opposite results. These findings suggest that specific characteristics of the video, such as content, length, and complexity, can significantly impact the performance of video summarization algorithms. It is crucial that future research considers these factors when evaluating and developing video summarization algorithms. Furthermore, our results highlight the need for a comprehensive and systematic evaluation of video summarization algorithms using diverse datasets

and scenarios, rather than relying on a single evaluation metric or scenario. By doing so, we can gain a better understanding of the strengths and limitations of different algorithms and guide their development for a wide range of applications.

Among other things, the video quality indicators created by the Laboratory for Image and Video Engineering (LIVE) team of the University of Texas at Austin (e.g., the FRIQUEE indicator and OG-IQA), the HOSA indicator, the WaDIQaM indicator, or the CORNIA indicator [10]. Given that it adds additional quality indicators to the collection of tools supporting the summarizing system, the planned research is justified in light of the experiment's objectives. Additional indicators could result in a summarizing system that generates summaries of greater quality (particularly in terms of choosing aesthetically appealing scenarios for the material recipient).

Our experimental results show that the proposed crowdsourcing technique outperforms other methods in terms of relevance and completeness. Our work contributes to filling this gap and examines the impact of the study's interface design on participants' perceptions of video quality. Overall, subjective quality evaluation of video summarizing algorithms may reveal their strengths and weaknesses. By gathering input from human assessors, we may discover areas for improvement and build stronger algorithms that can effectively capture the key information in the original video and display it in a compact and understandable manner.

6. CONCLUSION

In this paper, we proposed a new technique for video summarization using crowdsourcing. We designed an experiment in which individuals were hired to watch and summarize video clips [11]. We presented the effectiveness of the proposed technique by comparing it with other existing methods for video summarization. The results of our experiment showed that the proposed technique provides a significant improvement over other methods and is a promising approach for summarizing video content. The proposed technique has several advantages, such as being flexible, scalable, and cost-effective. The evaluation of video summarization algorithms is crucial, wherein subjective quality assessment plays a significant role. The aforementioned statement offers valuable insights into the user's perception of the summary and can be utilized to pinpoint potential areas for enhancement. The selection of the approach for evaluating subjective quality is contingent upon multiple factors and warrants meticulous consideration to guarantee the accuracy and consistency of the outcomes. Future work includes exploring different segmentation and selection methods and evaluating the proposed technique on larger datasets.

The results of the experiment suggest that crowdsourcing can be a useful tool for video summarization, as it allows for the aggregation of diverse opinions and perspectives. It outperforms other methods, with the quality of the summaries varying depending on the scenario. The crowdsourcing technique is an inexpensive and efficient method for summarizing video clips. Ultimately, the use of crowdsourcing for video summarization has the potential to revolutionize the way we create and consume video content, making it more comprehensive, diverse, and accessible for all.

The experiment has limitations, such as the limited number of scenarios. Future research could explore the use of more scenarios and the evaluation of the summaries by end-users to assess their usability and effectiveness.

LITERATURA

[1] T. Hoßfeld et al., "Best Practices for QoE Crowdsourcing: QoE Assessment With Crowdsourcing," in IEEE Transactions on Multimedia, vol. 16, no. 2, pp. 541-558, Feb. 2014. https://doi.org/10.1109/TMM.2013.2291663

[2] Wei-Tek Tsai, Li Zhang, Shufeng Hu, Zizheng Fan, Qianyu Wang, "Crowdtesting Practices and Models: An Empirical Approach," Information and Software Technology, Volume 154, 2023, 107103, ISSN 0950-5849. https://doi.org/10.1016/j.infsof.2022.107103

[3] M. Shahid et al., "Crowdsourcing-based subjective quality assessment of adaptive video streaming," 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX), Singapore, 2014, pp. 53-54. https://doi.org/10.1109/QoMEX.2014.6982289

[4] https://scenedetect.com/en/latest (Accessed on 19 April 2023)

[5] Romaniak, Piotr; Janowski, Lucjan; Leszczuk, Mikołaj; Papir, Zdzisław. "Perceptual quality assessment for H.264/AVC compression," In: 2012 IEEE Consumer Communications and Networking Conference (CCNC), pp. 597–602, IEEE, 2012. https://doi.org/10.1109/CCNC.2012.6181021

[6] https://qoe.agh.edu.pl/indicators/#indicators (Accessed on 20 April 2023)

[7] Leszczuk, Mikołaj; Kobosko, Marek; Nawała, Jakub; Korus, Filip; and Grega, Michał; 2023. "“In the Wild” Video Content as a Special Case of User Generated Content and a System for Its Recognition" Sensors 23, no. 4: 1769. https://doi.org/10.3390/s23041769

[8] Badiola, Aritz, Amaia Méndez Zorrilla, Begonya Garcia-Zapirain Soto, Michał Grega, Mikołaj Leszczuk, and Kamel Smaïli. “Evaluation of Improved Components of AMIS Project for Speech Recognition, Machine Translation and Video/Audio/Text Summarization.” Multimedia Communications, Services and Security, 2020, 320–31. https://doi.org/10.1007/978-3-030-59000-0_24

[9] Borchert, K., Seufert, A., Gamboa, E. et al. "In vitro vs in vivo: Does the study's interface design influence crowdsourced video QoE?," Qual User Exp 6, 1 (2021). https://doi.org/10.1007/s41233-020-00041-2

[10] Leszczuk, Mikołaj; Janowski, Lucjan; Nawała, Jakub; and Boev, Atanas; 2022. "Objective Video Quality Assessment Method for Face Recognition Tasks" Electronics 11, no. 8: 1167. https://doi.org/10.3390/electronics11081167

[11] A. Zlatintsi, P. Koutras, N. Efthymiou, P. Maragos, A. Potamianos and K. Pastra, "Quality evaluation of computational models for movie summarization," 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX), Pilos, Greece, 2015, pp. 1-6. https://doi.org/10.1109/QoMEX.2015.7148146

Algorytm ukrywania danych w obrazie oparty na klasyfikacji do wykrywania i rozpoznawania ukrytych informacji

Image data hiding algorithm based on classification for detecting and recognising hidden information

Paweł Duszejko[1]; Zbigniew Piotrowski[2]

[1] Wojskowa Akademia Techniczna, Warszawa, pawel.duszejko@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Warszawa, zbigniew.piotrowski@wat.edu.pl

DOI: 10.15199/59.2023.4.76

Streszczenie: **W artykule przedstawiono innowacyjny algorytm ukrywania danych w obrazach oparty na klasyfikacji. Metoda pozwala na niezauważalne dla ludzkiego oka ukrycie danych, a jednocześnie zapewnia możliwość ich późniejszego wykrycia i rozpoznania bez konieczności posiadania klucza lub oryginalnego obrazu. Przeprowadzone eksperymenty potwierdzają skuteczność i niezawodność tej metody w porównaniu z innymi algorytmami. Proponowane rozwiązanie może się przyczynić do rozwoju takich dziedzin jak: bezpieczeństwo informacyjne, ochrona prywatności, autoryzacja i znakowanie wodne.**

Abstract: **The article presents an innovative algorithm for data hiding in images based on classification. The method allows imperceptible data hiding to the human eye while also providing the ability to later detect and recognise the hidden data without needing a key or the original image. The conducted experiments confirm the effectiveness and reliability of this method compared to other algorithms. The proposed solution can contribute to developing fields such as information security, privacy protection, authentication, and watermarking.**

Słowa kluczowe: **sieci neuronowe, ukrywanie danych, watermarking, steganografia, klasyfikator**

Keywords: **neural networks, data hiding, watermarking, steganography, classifier**

1. WSTĘP

W dzisiejszych czasach obraz stał się jednym z najważniejszych mediów używanych zarówno w celach rozrywkowych, jak i biznesowych. W związku z tym jego ochrona przed nieautoryzowanym dostępem stała się priorytetem dla wielu przedsiębiorstw i osób prywatnych. Jednym z rozwiązań pozwalających na taką ochronę jest ukryty identyfikator osadzany w sposób, który nie wpływa na jakość wizualną. Algorytm ukrywania danych (ang. data hiding) oparty na klasyfikacji pozwala na skuteczne wykrycie i odszyfrowanie ukrytych danych w obrazie poprzez estymację skrytego identyfikatora klasy. Rozwiązanie to może stać się narzędziem znakowania wodnego, które będzie pozwalało na ochronę treści multimedialnych przed nieautoryzowanym dostępem.

Aktualnie wiodącym sposobem klasyfikacji obrazów jest wykorzystanie sieci neuronowych. Jedną z najpopularniejszych i najefektywniejszych jest ResNet [1]. Ta zaawansowana architektura została opracowana przez zespół Microsoft Research w 2015 roku. Składa się ona z wielu bloków, tak zwanych bloków rezydualnych (ang. residual blocks), z których każdy zawiera wiele warstw konwolucyjnych, normalizacji wsadowej i aktywacji ReLU. Jedną z cech wyróżniających ResNet spośród innych sieci neuronowych jest sposób przekazywania informacji między warstwami. Wykorzystuje tak zwane połączenia pomijające lub skokowe (ang. skip connections), które pozwalają na przekazywanie informacji między warstwami bezpośrednio z jednej do kilku kolejnych. Dzięki temu sieć ta jest w stanie nauczyć się reprezentacji reszty danych, a nie tylko informacji wprowadzanej przez kolejne warstwy. Umożliwia to budowanie bardzo głębokich sieci neuronowych, co wpływa pozytywnie na dokładność klasyfikacji obrazów. ResNet jest zwykle wstępnie szkolony na dużych zbiorach danych, takich jak ImageNet, co skutkuje dokładnym rozpoznawaniem obiektów i ich klas.

Autorzy postanowili sprawdzić, jak zachowa się architektura ResNet, gdy zostanie wykorzystana do klasyfikacji ukrytych danych. Do tego celu użyto algorytmu ukrywania danych w obrazie, który zapewnia wysoką niezauważalność osadzonej informacji. Zastosowano technikę treningu bez nadzoru, aby zidentyfikować i wydobyć cechy, które będą reprezentacją skrytych klas. Następnie wykorzystano przeszkoloną sieć ResNet do klasyfikacji obrazów na bazie ukrytych danych. Badanie pokazało, że ResNet można przystosować do tego celu, co otwiera nowe możliwości zastosowania sieci neuronowych w dziedzinie ukrywania danych.

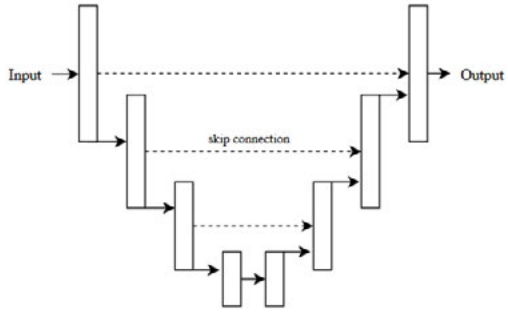
2. METODA

Opisywana metoda implementuje klasyczną doktrynę ukrywania danych opartą na koncepcji enkoder-dekoder. Jako nośnik informacji (ang. cover) wykorzystano obraz, a jako ukrytą wiadomość – skrytą reprezentację klasy wypracowaną przez enkoder, która następnie zostaje nałożona na nośnik. Dekoder jest w stanie odczytać ukrytą informację z obrazu i na tej podstawie dokonać estymacji identyfikatora klasy.

2.1 Opis architektury

Realizacja enkodera oparta jest na architekturze sieci U-Net [2], której schemat ideowy pokazany został na *Rys. 1*. Sieć U-Net pierwotnie została zaprojektowana do rozwiązywania problemu segmentacji obrazów. W

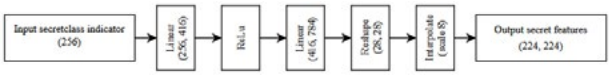
tym przypadku autorzy wykorzystali ją do osadzenia skrytej informacji. Blok enkodera tej sieci to seria warstw konwolucyjnych, które służą do ekstrakcji cech z obrazu wejściowego. Warstwy te zmniejszają rozmiar obrazu, jednocześnie zwiększając liczbę cech tworząc na każdym etapie mapy cech (ang. feature maps).



Rys. 1. Schemat sieci U-Net

Blok dekodera jest serią warstw transponowanych konwolucji, które służą do odtworzenia obrazu wejściowego na podstawie map cech otrzymanych z bloków kodera. W trakcie działania bloków dekodera rozmiar obrazu stopniowo wzrasta, jednocześnie redukując liczbę cech. Architekturę tę wyróżnia zastosowanie, podobnie jak w przypadku ResNet połączeń pomijalnych. Dzięki temu informacja o danych wejściowych jest zachowywana na wyższych poziomach sieci, umożliwiając przechowywanie cech lokalnej struktury obrazu i wykorzystanie ich przy przetwarzaniu na wyższych poziomach abstrakcji. Sprzyja to analizie kontekstu lokalnego i globalnego w obrazie. W głębokich sieciach neuronowych powszechny jest problem zanikających i eksplodujących gradientów. W pierwszym przypadku w procesie propagacji wstecznej gradienty maleją co utrudnia skuteczną aktualizację wag sieci. W drugim przypadku następuje gwałtowny wzrost gradientów, co może prowadzić do niestabilności i znikomej zbieżności algorytmu uczącego. Połączenia skokowe pozwalają na uniknięcie tych problemów, ponieważ umożliwiają bezpośredni przepływ informacji z wyższych warstw do niższych.

W celu przekazania sieci U-Net wybranego identyfikatora skrytej klasy, konieczne jest jego wstępne przygotowanie. Zaproponowano osobny moduł do realizacji tego zadania. Nazwano go siecią przygotowującą wiadomość skrytą. *Rys. 2*. przedstawia architekturę modułu.



Rys. 2. Schemat sieci przetwarzającej identyfikator skrytej klasy

Sieć ta ma za zadanie ekstrahować cechy identyfikatora skrytej klasy oraz sprowadzić je do rozmiaru obrazu nośnego, umożliwiając tym samym konkatenacje tych danych w późniejszym czasie.

Ostatnim elementem systemu jest wspomniany już klasyfikator skrytej sygnatury. Oparty został na popularnej sieci ResNet w wersji 34-blokowej (ResNet-34). Architektura sieci została przedstawiona na *Rys. 3*.

Rys. 3. Architektura sieci ResNet-34 [https://huggingface.co/microsoft/resnet-34]

Jednym z głównych celów omawianego algorytmu była prezentacja możliwości przystosowania sieci ResNet do klasyfikacji skrytych artefaktów. Zmienia to drastycznie zachowanie sieci, która w swojej natywnej formie uczyła się klasyfikować obiekty ze świata rzeczywistego poprzez ekstrakcję ich cech wizualnych, co nie było by możliwe w przypadku skrytych artefaktów.
W modelu zdefiniowano 255 klas bazowych, które reprezentują klasy skrytych znaczników na obrazach. Jednakże, aby zapobiec błędnym odczytom i pomyłkom, została stworzona 256-a klasa – dyskryminator. Wskazuje on, że dane zdjęcie nie posiada skrytych artefaktów i może być próbą oszukania algorytmu. Może on rozpoznawać również zbyt uszkodzony nośnik. Ta innowacyjna funkcjonalność pozwala na uniknięcie błędnych wskazań i jest nowością względem innych algorytmów.

2.2 Opis przepływu danych

Pierwszym krokiem w potoku przetwarzania danych (ang. pipeline) jest użycie sieci przygotowującej identyfikator skrytej klasy, poprzez rozszerzenie go do obiektu o rozmiarze nośnika (224x224), który reprezentuje 1-kanałową mapę cech. Następnie mapa cech jest konkatenowana z obrazem nośnym, który jest 3-kanałowy (format RGB) i ma ten sam rozmiar. Wynikowo otrzymano 4-kanałową strukturę 224x224, jako wejście dla sieci U-Net – wejście do enkodera. Na wyjściu enkodera otrzymana struktura posiada 3-kanały o rozmiarze 224x224. Jest to już reprezentacja obrazu z nałożoną sygnaturą – produktem enkodera systemu, dalej nazywanym stegano-obiektem.

W przypadku części dekodera systemu zastosowano transfer learning, czyli wykorzystanie już wytrenowanych wag z sieci ResNet, która nauczyła się klasyfikacji 1000 klas obiektów na zbiorze danych ImageNet. Zmniejszono rozmiar warstwy klasyfikacyjnej z 1000 do 256, aby dostosować ją do potrzeb omawianego algorytmu. Na wyjściu dekodera otrzymywana jest informacja o prawdopodobieństwie przynależności stegano-obiektu do każdej z klas.
Wszystkie trzy sieci: przygotowująca identyfikator klasy, U-Net oraz ResNet tworzą kompletną sekwencję przetwarzania danych – system przedstawiono na *Rys. 4*.

Rys. 4. Graf przepływu danych w proponowanym systemie

Istotnym elementem systemu jest blok przekształceń obrazu przedstawiony w postaci żółtego bloku na *Rys. 4*. Ma on za zadanie uodpornić algorytm na zagrożenia związane z przesyłem stegano-obiektu przez tor transmisyjny.

3. WYNIKI EKPERYMENTÓW

W tym rozdziale omówione zostaną rezultaty eksperymentu z wykorzystaniem architektury opisanej w poprzednich sekcjach.

3.1 Parametry treningu

Do eksperymentu wykorzystano zbiór danych *Hollywood2 Human Actions and Scenes Dataset [3]* zawierający ponad 800 fragmentów filmów różnych gatunków. Filmy te zostały poddane procesowi próbkowania. Zapisywano co 35 klatkę filmu, co pozwoliło na uzyskanie około 40 tysięcy zdjęć treningowych w różnych rozdzielczościach. Do etapu treningu sieci losowano fragmenty o rozmiarze 224x224 pikseli. Zapewniało to wysokie zróżnicowanie próbek podczas treningu i zapobiegało przeuczeniu sieci oraz monotonii danych wejściowych.

W proponowanym systemie występują dwie główne funkcje straty. Pierwsza z nich *secret class loss* skupia się na poprawnej klasyfikacji stegano-obiektów i wykorzystuje entropie krzyżową do realizacji tego zadania. Funkcja ta jest związana z siecią dekodera ResNet. Druga funkcja straty *image loss* jest odpowiedzialna za utrzymanie jak najwyższej transparentności obrazu po osadzeniu skrytych artefaktów. Do tego celu wykorzystywano trzy miary podobieństwa: SSIM (ang. structural similarity index), MSE (ang. mean squared error) oraz LPIPS (ang. learned perceptual similarity metric) [4]. Funkcja ta dotyczyła enkodera, czyli sieci U-Net.

Parametr SSIM opisany jest wzorem:

$$SSIM = \frac{(2\mu_x\mu_y + k_1R)(2\sigma_{xy} + k_2R)}{(\mu_x^2 + \mu_y^2 + k_1R)(\sigma_x^2 + \sigma_y^2 + k_2R)}$$

gdzie:
$\mu_x, \mu_y - średnia\ obrazu\ x\ i\ y$
$\sigma_x^2, \sigma_y^2 - wariancja\ obrazu\ x\ i\ y$

Parametr MSE opisany jest wzorem:

$$MSE = \frac{1}{WH}\sum_{i=1}^{H}\sum_{j=1}^{W}(X_{i,j} - Y_{i,j})^2$$

gdzie:
$W - szerokość\ obrazu$
$H - wysokość\ obrazu$

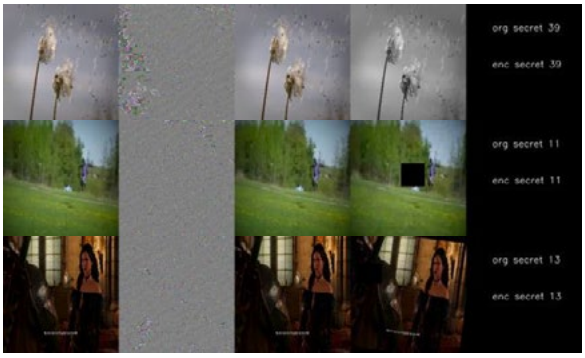
Obie funkcje straty były sumowane i wykorzystywane do propagacji wstecznej (ang. backpropagation). Zabieg ten umożliwia trenowanie całego modelu.

Dodatkowym elementem potoku treningowego była wspomniana sekcja szumów z *Rys. 4*, która obejmowała kilka operacji przetwarzania obrazów. Pierwszą z nich była kompresja JPEG, która symulowała straty danych wynikające z kompresji plików. Następnie dodawano szum gaussowski oraz rozmycie, które miały na celu zwiększenie odporności modelu na zakłócenia w torze nadawczo-odbiorczym, występujące w zależności od jakości transmisji. Ostatecznie wykonywano operację zakrywania poprzez losowe usuwanie fragmentów obrazów, aby zwiększyć odporność modelu na celowe ataki.

Wprowadzenie tych przekształceń miało za zadanie nauczyć koder jak efektywnie i odpornie osadzać dane, a klasyfikator generalizacji skrytych danych. Pozwalało to również uniknąć nadmiernego dopasowania do danych treningowych.
Dzięki tym działaniom model był w stanie lepiej radzić sobie z zakłóceniami w danych oraz osiągał wyższą skuteczność w zadaniu klasyfikacji i osadzania.

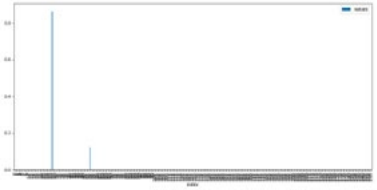
3.2 Otrzymane rezultaty

W eksperymencie model trenowano przez 300 epok i mierzono wartość sumarycznej funkcji straty. Ewaluacja odbywała się na zbiorze testowym. Składał się on z 1000 elementów, które nie były wykorzystane w procesie treningu i nie były znane sieci. W ten sposób można było ocenić ogólną skuteczność modelu w klasyfikacji skrytych artefaktów. Poglądowe rezultaty wizualne zostały zgromadzone i przedstawione na Rys. 5. Patrząc od lewej strony obrazu przedstawiają one kolejno: obraz nośny, nanoszone artefakty pomnożone przez wartość 255, obraz z nałożonym znakiem wodnym, obraz z nałożonym znakiem wodnym z dodanymi zakłóceniami oraz wyniki klasyfikacji.



Rys. 5. Otrzymane rezultaty ewaluacji sieci

Wybór klasy odbywa się poprzez wyznaczenie instancji z największym prawdopodobieństwem obliczonym przez funkcję softmax – przykładowy graf dla jednego z obrazów ze zbioru testowego przedstawiono na Rys. 6. Prawdopodobieństwo jest zdecydowanie najwyższe dla jednej klasy. Jego wartość może być traktowana jako rodzaj wskaźnika zaufania (ang. confidence index).



Rys. 6. Prawdopodobieństwo dla klas w systemie wyznaczone za pomocą funkcji softmax

Na bazie wspomnianych danych testowych określono metryki statystyczne, uwzględniając również parametr PSNR (ang. peak signal-to-noise ratio):

Dla transparentności obrazu:

- średnia wartość SSIM: 0.9985
- średnia wartość PSNR: 54.45
- średnia wartość MSE: 0.0000036

Dla skuteczności klasyfikacji:

- Accuracy: 0.77
- F1 score: 0.84

Dokonano również zestawienia otrzymanych metryk w stosunku do innych najbardziej rozpoznawalnych prac w omawianej dziedzinie. Zilustrowano je w Tab. 1.

Tab. 1. Porównanie kluczowych parametrów wybranych metod

Nazwa rozwiązania	*SSIM, PSNR*	*Pojemność*	*Szumy*
SteganoGan [5]	0,925, 40,27	768 000b	Nie
HiDDeN [6]	-, 37,27	52b	Tak
SteganoCNN [7]	0.981, 35.852	130 072b	Nie
Omawiana metoda	0.9985, 54,45	8b	Tak

Warto zauważyć, że w porównaniu do innych rozwiązań, zaproponowana metoda nie opiera się na regresji skrytej wiadomości, a na klasyfikacji. W związku z czym kolumna "pojemność" ma stosunkowo niskie wartości. Jednakże jest to rekompensowane przez wysoką transparentność, co jest kluczowym parametrem w wielu zastosowaniach.

4. PODSUMOWANIE

Podsumowując omawiana metoda jest innowacyjnym podejściem do zagadnienia data hiding, polegającym na rozpoznawaniu ukrytych klas. Zapewniono wysoką transparentność i odporność produktu algorytmu na zakłócenia w torze nadawczo-odbiorczym. W ramach ewaluacji na zbiorze testowym liczącym 1000 zdjęć osiągnięto zadowalające wyniki z metryką F1 wynoszącą powyżej 0.8. Warto również zaznaczyć, że zaprojektowano 256 klas, w tym jedną służącą jako dyskryminator, co pozwala na uniknięcie błędów wynikających z braku ukrytych danych lub zbyt uszkodzonego nośnika.

LITERATURA

[1] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. "Deep Residual Learning for Image Recognition." ArXiv:1512.03385.

[2] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation." ArXiv:1505.04597

[3] Marszałek, Marcin, Ivan Laptev, and Cordelia Schmid. 2009. "Actions in Context." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[4] Zhang, Richard, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[5] Zhang, Kevin Alex, Alfredo Cuesta-Infante, Lei Xu, and Kalyan Veeramachaneni. 2019. "SteganoGAN: High Capacity Image Steganography with GANs." ArXiv:1901.03892

[6] Zhu, Jiren, Russell Kaplan, Justin Johnson, and Li Fei-Fei. 2018. "HiDDeN: Hiding Data With Deep Networks." ArXiv:1807.09937 [cs.CV]

[7] Duan, Xintao, Nao Liu, Mengxiao Gou, Wenxin Wang, and Chuan Qin. 2020. "SteganoCNN: Image Steganography with Generalization Ability

ZWIĘKSZANIE POTENCJAŁU SZTUCZNEJ INTELIGENCJI GENERATYWNEJ DZIĘKI PROMPT ENGINEERING

INCREASING THE POTENTIAL OF GENERATIVE ARTIFICIAL INTELLIGENCE WITH PROMPT ENGINEERING

Marta Bistroń[1]; Zbigniew Piotrowski[2]

[1] Wojskowa Akademia Techniczna, Warszawa, marta.bistron@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Warszawa, zbigniew.piotrowski@wat.edu.pl

DOI: 10.15199/59.2023.4.77

Streszczenie: **W artykule opisano różne techniki stosowane w algorytmach sztucznej inteligencji generatywnej, takie jak modele oparte na rozkładach prawdopodobieństwa, modele wariancyjne oraz modele sekwencyjne. Wyjaśniono podstawy tych technik oraz omówiono ich zastosowania w generowaniu obrazów, muzyki, tekstu czy mowy. Artykuł podkreśla znaczenie algorytmów generative AI jako narzędzi do twórczego generowania treści oraz prezentuje możliwe sposoby zwiększenia efektywności generowania tych treści z wykorzystaniem techniki prompt engineering.**

Abstract: **The article describes various techniques used in generative artificial intelligence algorithms, such as models based on probability distributions, variance models and sequential models. The basics of these techniques are explained and their applications in generating images, music, text or speech are discussed. The article emphasizes the importance of generative AI algorithms as tools for creative content generation and presents possible ways to increase the efficiency of generating this content using the prompt engineering technique.**

Słowa kluczowe: **autoenkoder, model generatywny, prompt engineering, sztuczna inteligencja generatywna, transformer.**

Keywords: **autoencoder, generative artificial intelligence, generative model, prompt engineering, transformer.**

1. WSTĘP

Sztuczna inteligencja generatywna (Generative Artificial Intelligence, Generative AI) jest to dziedzina sztucznej inteligencji, która zajmuje się opracowywaniem oraz implementacją modeli i systemów generatywnych zdolnych do tworzenia nowych, autentycznych i nieznanych wcześniej danych na podstawie wzorców i reguł przyswojonych z istniejącego zbioru danych treningowych. Głównym zadaniem Generative AI jest tworzenie nowych danych, które dla użytkownika wydają się autentyczne i są trudne do odróżnienia od danych rzeczywistych. Celem tego typu algorytmów jest tworzenie wysoce realistycznych i złożonych treści, które imitują ludzką kreatywność, co pozwala na zastosowanie modeli generatywnych w wielu różnych branżach i produktach [1].

W ostatnim czasie termin generatywnej sztucznej inteligencji stał się bardzo popularny medialnie, a dzięki temu powszechniejszy wśród społeczeństwa, ze względu na rosnącą popularność aplikacji bazujących na modelach generatywnej sztucznej inteligencji opracowanych przez firmę OpenAI – ChatGPT (konwersacyjny chatbot) [2] oraz DALL-E (generator obrazów typu text-to-image) [3]. Obydwa rozwiązania bazują na architekturze GPT (Generative Pre-Trained Transformer). Jest to model językowy bazujący na architekturze sieci neuronowych typu transformer – sieci sekwencyjnych, odznaczających się wysoką efektywnością w problemach przetwarzania języka naturalnego. Model GPT jest trenowany w trybie „pre-training" na bardzo obszernych zbiorach danych z Internetu, a następnie jest dostosowywany do rozwiązywania konkretnych problemów z różnych obszarów.

Modele generatywne cieszą się tak dużą popularnością właśnie ze względu na możliwość generowania nowych treści różnego typu, dostosowanych do konkretnych zastosowań – tworzeniu kodu komputerowego [4], pisania esejów, e-maili [5], podpisów w mediach społecznościowych, obrazów, wierszy, muzyki [6], czy też formuł programu Excel. Dzięki temu zainteresowanie sztuczną inteligencją generatywną ciągle rośnie, ponieważ coraz więcej firm dostrzega jej potencjał i znajduje nowe zastosowania dla algorytmów. Szczególnie rozwojową branżą w tym obszarze jest marketing, generatywna AI może być wykorzystywana do tworzenia nowych treści reklamowych dostosowanych do odbiorców na podstawie rekomendacji [7], interesujących materiałów promocyjnych, ofert, a nawet kanałów komunikacji dopasowanych do indywidualnych potrzeb i zainteresowań. Sztuczna inteligencja może też wspierać proces prototypowania i tworzenia nowych produktów poprzez generowanie wizualizacji nowych produktów, projektowanie opakowań, czy też tworzenie interaktywnych demonstratorów.

Skuteczność i efektywność opisanych powyżej zastosowań sztucznej inteligencji generatywnej jest kwestią bardzo złożoną i nie zależy tylko od procesu treningu algorytmu i zastosowanych do niego danych. Kluczowe są dane wejściowe dostarczane do algForytmu już na etapie jego użytkowania, tzw. podpowiedzi, wskazówki, instrukcję (z ang. *prompt*). Jest to zestaw instrukcji lub parametrów, który informuje narzędzie w jaki sposób powinny zostać wygenerowane dane wyjściowe. W zależności od rodzaju modelu może to być proste pytanie do chatbota lub sparametryzowana informacja, pozwalająca uzyskać bardziej spersonalizowane treści. Pisanie podpowiedzi umożliwiających uzyskanie satysfakcjonujących odpowiedzi może wymagać wielu prób i odpowiedniego doświadczenia w pracy z danym modelem, czego skutkiem jest nowa, dynamicznie rozwijająca się praktyka *prompt engineering* (inżynieria promptów). Jest to proces dostosowywania podpowiedzi, tak aby zmaksymalizować trafność uzyskanej odpowiedzi do oczekiwań odbiorcy,

czyli spełnienie określonych kryteriów, takich jak np.: precyzja, spójność, styl wypowiedzi, bezstronność, czy też zgodność z pewnymi wartościami lub wzorcami.

Dalsza cześć artykułu została ułożona w następujący sposób: w rozdziale 2 dokonano podziału algorytmów sztucznej inteligencji generatywnej oraz opisano ich architekturę i zasadę działania, w rozdziale 3 omówiono koncepcję inżynierii promptów oraz opisano najważniejsze reguły jego wykorzystania dla wybranych narzędzi, w rozdziale 4 podsumowano wyniki rozważań.

2. ALGORYTMY SZTUCZNEJ INTELIGENCJI GENERATYWNEJ

Sztuczna inteligencja generatywna może być wykorzystywana w wielu różnych obszarach dotyczących zarówno działalności artystycznej związanej z generowaniem nowych treści, jak i w zastosowaniach praktycznych. W ramach sztucznej inteligencji generatywnej stosowanych jest wiele technik i podejść, jednak można wyróżnić 3 główne i najpopularniejsze techniki:

- generatywne modele probabilistyczne,
- modele generatywne oparte na kodowaniu i dekodowaniu,
- modele oparte na sekwencjach.

2.1. Generatywne modele probabilistyczne

Generatywne modele probabilistyczne to technika uczenia nienadzorowanego, która opiera się na probabilistycznych rozkładach danych, modele uczą się generować nowe dane o rozkładzie prawdopodobieństwa zbliżonym do rozkładu danych uczących. W praktyce najczęściej wykorzystywane są probabilistyczne modele graficzne (*Probablistic Graphical Models, PGM*) [8]. Wśród takich rozwiązań najpopularniejsze są aktualnie modele przeciwstawne – tzw. sieci GAN (*Generative Adversarial Networks*) [9]. GAN nie jest typowym modelem generatywnym opartym na prawdopodobieństwie, takim jak modele generatywne oparte na sieciach Boltzmanna lub sieciach Bayesowskich. GAN jest oparty na podejściu adwersyjnym, składa się z dwóch komponentów: generatora i dyskryminatora. Generator generuje nowe przykłady na podstawie datasetu, natomiast dyskryminator stara się odróżnić prawdziwe dane od danych wygenerowanych przez generator. Oba komponenty nieustannie konkurują ze sobą - generator uczy się generować wiarygodne przykłady, a dyskryminator uczy się odróżniać dane prawdziwe od podróbek (*fake*).

Innym podejściem są modele autoregresyjne, jak np. PixelRNN [10] lub PixelCNN [11]. Podstawowa różnica polega na tym, że sieci GAN uczą się dystrybucji niejawnej (*implicit*), a modele autoregresyjne dystrybucji jawnej (*explicit*). W przypadku danych jawnie dystrybuowanych każdy punkt danych lub jego etykieta jest jawny i dostępny dla modelu, można bezpośrednio odwołać się do konkretnego punktu danych. Dzięki temu modele te w przeciwieństwie do GAN zwracają wartość gęstości prawdopodobieństwa, a ich trening jest bardziej stabilny [12].

2.2. Modele oparte na kodowaniu i dekodowaniu

Modele generatywne oparte na kodowaniu i dekodowaniu wykorzystują architekturę autoenkodera, czyli współpracujących ze sobą kodera i dekodera, w celu przekształcenia pewnej reprezentacji wejściowej do pożądanej reprezentacji wyjściowej. Przykładem takiego algorytmu jest wariancyjny autoenkoder (*Variational Autoencoder, VAE*), który generuje nowe dane poprzez zakodowanie ich przez enkoder w tzw. przestrzeni latentnej (przestrzeni reprezentacji), a następnie odtworzenie danych z przestrzeni latentnej przez dekoder [13].

Poza standardowym VAE wyróżnia się również inne warianty dostosowane do konkretnych problemów. Jednym z rozwiązań jest czynnikowy autoenkoder *(Factor VAE*), który warunkuje rozkład danych w warstwie latentnej na pewne zmienne dyskretne, co ma na celu wyizolowanie różnych czynników, które mogą wpływać na dane wejściowe. Wpłynie to na skuteczność działania algorytmu w przypadku analizy przyczynowo skutkowej danych lub podczas redukcji wymiarowości [14]. Innym wariantem jest adwersyjny autoenkoder (*Adersarial Autoencoder, AAE*), który łączy w sobie elementy tradycyjnego autoenkodera oraz sieci GAN. Architektura składa się z trzech komponentów: enkodera, dekodera oraz dyskryminatora. Dyskryminator uczy się odróżniać prawdziwe przykłady od generowanych przez dekoder, natomiast enkoder i dekoder próbują wygenerować jak najbardziej realne dane [15]. Ostatnim popularnym podejściem jest hierarchiczny autoenkoder (*Hierarchical VAE*), który wprowadza hierarchię do przestrzeni latentnej, co pozwala tworzyć bardziej złożone reprezentacje danych. Przestrzeń latentna jest podzielona na warstwy, z których każda reprezentuje różne poziomy abstrakcji danych. Dzięki temu model może generować bardziej złożone struktury, takie jak np.: obiekty składające się z części składowych [16].

2.3. Modele sekwencyjne

Modele oparte na sekwencjach uczą się zależności pomiędzy elementami sekwencji, tak aby generować nowe sekwencje na podstawie wzorców rozpoznanych podczas przetwarzania zbioru danych treningowych. Dane sekwencyjne mogą obejmować strumienie tekstu, strumienie dźwięku, wideo lub inne dowolne szeregi czasowe. Pierwszymi modelami sekwencyjnymi były rekurencyjne sieci neuronowe (*RNN*) [17]. Są one wyposażone w pamięć wewnętrzną, która pozwala przechowywać informacje o wcześniejszych sekwencjach czasowych podczas generowania kolejnych elementów sekwencji. Sieć ma charakter rekurencyjny, ponieważ wykonuje tę samą funkcję dla każdego wejścia danych. Wyróżnia się dwa podstawowe rodzaje sieci rekurencyjnych – *Long Short-Term Memory* (LSTM) oraz *Gated Recurrent Unit* (GRU) [18]. Sieci LSTM mają więcej bramek i parametrów niż sieci GRU, co pozwala na większą elastyczność oraz wydajność, ale powoduje też większy koszt obliczeniowy oraz zwiększa ryzyka przeuczenia.

Aktualnie do przetwarzania sekwencji wykorzystuje się głównie modele typu transformer [19]. Główną różnicą w stosunku do tradycyjnych modeli sekwencyjnych jest zastosowanie mechanizmu samoatencji, do bardziej efektywnego wyszukiwania zależności pomiędzy róż-

nymi elementami sekwencji danych, co można sprowadzić do odpowiedzi na pytanie, na której części sekwencji wejściowej powinien skupić się model. Każde wejście do modułu samoatencji posiada trzy reprezentacje: klucz (*key*), zapytanie (*query*) oraz wartość (*value*), ich wartości są wyznaczane na podstawie macierzy wag, które na początku są najczęściej inicjalizowane losowo. Następnie wyznaczana jest wartość współczynnika uwagi (*attention score*), w tym celu należy wyznaczyć wartość iloczynu skalarnego pomiędzy macierzą zapytania danego wejścia a wszystkimi macierzami kluczy. Uzyskany wynik jest poddawany działaniu funkcji softmax. Dla każdego wejścia współczynnik uwagi jest wymnażany z macierzą wartości, zsumowanie uzyskanych macierzy pozwala uzyskać wynik końcowy (wartość wyjściową) dla danego wejścia. Opisany mechanizm samoatencji pozwala rozwiązać problem zanikających gradientów, co było częstym problem podczas analizy długich sekwencji z wykorzystaniem klasycznych sieci rekurencyjnych. W transformerach po każdej warstwie samoatencji wykorzystuje się warstwy liniowe (*fully connected layers*), które pomagają w dalszym modelowaniu zależności między elementami sekwencji.

Przykładem transformera jest popularny model językowy GPT [20], charakteryzujący się wysoką efektywnością w przetwarzaniu języka naturalnego. Cechą charakterystyczną modelu jest zastosowanie mechanizmu generacji autoregresyjnej. Oznacza to, że GPT sekwencyjnie generuje kolejne słowa na podstawie wcześniej wygenerowanego tekstu, co pozwala na uwzględnienie kontekstu całej wypowiedzi oraz generowanie bardziej płynnego tekstu. Ponadto trening modelu obejmuje dwa etapy: pretraining, czyli naukę wzorców językowych na podstawie rozległych zbiorów danych oraz fine-tuning, czyli dotrenowanie modelu do rozwiązywania konkretnego problemu jak np.: tłumaczenie maszynowe lub odpowiadanie na pytania użytkownika.

3. PROMPT ENGINEERING

W celu jak optymalnego wykorzystania potencjału modeli generatywnych konieczne jest zastosowanie *prompt enegineering*, czyli praktyki polegającej na optymalizacji wejściowych podpowiedzi tzw. promptów dla modeli językowych. Ma to duże znaczenie zwłaszcza w obszarze modeli językowych, które zazwyczaj są bardzo elastyczne i mogą generować różne odpowiedzi w zależności od kontekstu zapytania. Prompt engineering wymaga licznych eksperymentów i analiz ukierunkowanych na dany model. Nie jest to tradycyjny obszar programowania, jednak coraz częściej można spotkać się z opracowanymi w dowolnym języku narzędziami, które komunikują się bezpośrednio z API danego modelu np. GPT-3.5 oraz wysyłają odpowiednie zapytania i analizują odpowiedzi, w celu weryfikacji efektywności poszczególnych promptów.

Firma OpenAI opracowała tzw. Cookbook, w którym udostępniono przykładowy kod w języku Python umożliwiający wykonanie typowych zadań z wykorzystaniem interfejsu API Open AI [21]. Ponadto opracowany został również zbiór tzw. najlepszych praktyk dotyczących prompt engineering z wykorzystaniem API OpenAI [22]. Poradnik obejmuje 8 reguł: wykorzystywanie najnowszej dostępnej wersji modelu generatywnego; zamieszczanie instrukcji na początku promptu oraz oddzielanie instrukcji od kontekstu za pomocą cudzysłowów; podawanie jak najbardziej konkretnych i szczegółowych informacji dotyczących kontekstu, wyniku, formatu, stylu, czy długości wypowiedzi; wskazywanie oczekiwanego formatu danych wyjściowych na podstawie przykładu; podawanie przykładu efektu działania modelu; redukowanie nieprecyzyjnych, „poetyckich" opisów; zastępowanie instrukcji negatywnych pozytywnymi; stosowanie „słów kluczy" nakierowujących model na konkretny rodzaj działania.

Równie popularnym wykorzystaniem generative AI co chatboty konwersacyjne są tzw. modele *text-to-image*, które przyjmują na wejściu opis tekstowy, na podstawie którego generują wyjściową grafikę. Zasadnicza architektura tego typu modeli bazuje na sieciach GAN. Na wejście generatora podawany jest opis tekstowy tzw. adnotacja, opis jest sprowadzany do postaci wektorowej z wykorzystaniem modelu sekwencyjnego (np. transformer). Na tej podstawie generator bazujący głównie na sieciach konwolucyjnych generuje obraz odpowiadający opisowi. Podczas procesu uczenia model analizuje pary tekst-obraz, które są wykorzystywane do nauki odpowiednich powiązań i wzorców.

Do najpopularniejszych rozwiązań generatorów grafiki należą DALLE-2 opracowany przez OpenAI oraz Midjourney. W przypadku Midjourney dostępnych jest wiele poradników umożliwiających wygenerowanie danych najbardziej zbliżonych do oczekiwań. Zasadnicza idea tworzenia promptu dla aplikacji polega na podaniu trzech typów danych, które składają się na finalną podpowiedź: adresy URL przykładowych obrazów, które pomagają zdefiniować styl i treść wyniku końcowego; opis tekstowy zawierający informację co powinno zostać wygenerowane; parametry, które pozwalają zmienić sposób generowania obrazu [23]. Kluczową funkcję pełnią parametry, które pozwalają w prosty sposób modyfikować wygląd grafiki za pomocą słów kluczowych. Do podstawowych parametrów należy m.in.: opis stylu w ramach którego możliwe jest podanie nazwiska artysty, do którego twórczości ma nawiązywać grafika lub określenie właściwego sposobu renderowania. Wiele parametrów przyjmuje też wartość liczbową np.: poziom abstrakcji, proporcje obrazu, poziom podobieństwa do obrazu wzorcowego. Możliwe jest również określenie konkretnej rozdzielczości, czy też zdefiniowanie prostego parametru „no", który określa co nie powinno zostać wygenerowane jak element grafiki.

4. PODSUMOWANIE

Celem artykułu było zaprezentowanie wybranych rozwiązań sztucznej inteligencji generatywnej oraz uwypuklenie wpływu inżynierii promptów na efektywność jej działania. Sztuczna inteligencja generatywna cieszy się aktualnie coraz większym zainteresowaniem społecznym, co wynika również w dużej mierze z medialności zagadnienia. Modele generatywne umożliwiają generowanie tekstu, grafiki, muzyki, a nawet kodu źródłowego, co

może być wykorzystywane w różnego rodzaju aplikacjach o zastosowaniach rozrywkowych i komercyjnych. Pomimo pozornej łatwości korzystania z narzędzi typu ChatGPT, modele generatywne, w celu pełnego wykorzystania ich potencjału powinny być obsługiwane we właściwy sposób, tak aby zwracana informacja była pełna i odpowiadała intencji zapytania użytkownika. W kontekście rozważań prawnych dotyczących ewentualnych praw autorskich do efektów działania modelu generatywnego, jakość sformułowanego promptu, która wpływa na ostateczną formę wyniku, jest bardzo istotna. Prompt engineering, jako technika dostosowywania podpowiedzi do modelu generatywnego, jest stosunkowym nowym, rozwijającym się obszarem wiedzy i techniki. Prompt engineer, co można przetłumaczyć dosłownie jako inżynier podpowiedzi, to nowy zawód, który zaczyna powoli zmieniać branże IT i został już okrzyknięty najmodniejszym zawodem IT. Sam Altman, CEO OpenAI, napisał na swoim Twiterze, że "Napisanie naprawdę świetnej podpowiedzi dla chatbota jest wysoko cenioną umiejętnością i przykładem programowania z użyciem naturalnego języka".

LITERATURA

[1] Foster David. 2019. „Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play. First Edition". Beijing, O'Reilly.

[2] Radford Alec et al. 2019. "Language Models are Unsupervised Multitask Learners"

[3] OpenAI. 2021. „DALL·E: Creating images from text", accessed: 22.05.2023, https://openai.com/research/dall-e.

[4] Romero Alberto. 2021. "GitHub Copilot — A New Generation of AI Programmers". Towards Data Science, accessed: 22.05.2023, https://towardsdatascience.com/github-copilot-a-new-generation-of-ai-programmers-327e3c7ef3ae.

[5] Hiken Asa. 2023. "How generative AI is creating custom content in email marketing".AdAge, accessed: 22.05.2023, https://adage.com/article/digital-marketing-ad-tech-news/how-generative-ai-creating-custom-content-email-marketing/2475106.

[6] Kaliakatsos-Papakostas Maximos, Floros Andreas, Vrahatis Michael N.2020. "Artificial intelligence methods for music generation: a review and future perspectives" in *Nature-Inspired Computation and Swarm Intelligence*. Academic Press. 217-245.

[7] Deveau Richelle, Joseph Griffin Sonia, Reis Steve. 2023. „AI-powered marketing and sales reach new heights with generative AI". McKinsey & Company, accessed: 22.05.2023, https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/ai-powered-marketing-and-sales-reach-new-heights-with-generative-ai.

[8] Samiullah M., Albrecht D., Nicholson A. E., Ahmed, C. F. 2022. „A Review on Probabilistic Graphical Models and Tools". *Dhaka University Journal of Applied Science and Engineering*, 6(2), 82–93.

[9] Goodfellow Ian. et al. 2014. „Generative adversarial nets". *Advances in neural information processing systems*, 2672–2680.

[10] Van den Oord Aaron. et all. 2016. "Pixel Recurrent Neural Networks" *Proceedings of the 33rd International Conference on Machine Learning.*

[11] Van den Oord Aaron. et all. 2016. "Conditional Image Generation with PixelCNN Decoders*" 30th Conference on Neural Information Processing Systems*, Barcelona, Spain.

[12] Sharm Harashit. 2017. "Auto-Regressive Generative Models (PixelRNN, PixelCNN++)" Towards Data Science, accessed: 22.05.2023, https://towardsdatascience.com/auto-regressive-generative-models-pixelrnn-pixelcnn-32d192911173.

[13] Kingma Diederik P., Welling Max. 2019. "An Introduction to Variational Autoencoders" *Foundations and Trends in Machine Learning*: 12 (4), 307-392.

[14] Hyunjik Kim, Andriy Mnih. 2019. "Disentangling by Factorising". *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, PMLR 80, 2018

[15] Makhzani A., Shlens J., Jaitly N., Goodfellow I., Frey B. 2015. "Adversarial Autoencoders", arXiv:1511.05644v2.

[16] Vahdat Arash, Kautz Jan. 2020. "NVAE: A Deep Hierarchical Variational Autoencoder". *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada

[17] Sherstinsky Alex. et al. 2020. "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network". *Physica D: Nonlinear Phenomena*, 404, 132306.

[18] Chung J. et al. 2014. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". *NIPS 2014 Deep Learning and Representation Learning Workshop*

[19] Vaswani A. et al. 2017. "Attention is all you need". *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000-6010.

[20] Radford Alec, et al. 2018. "Improving Language Understanding by Generative Pre-Training".

[21] OpenAI. "OpenAI Cookbook", accessed: 22.05.2023, https://github.com/openai/openai-cookbook.

[22] Open AI. "Best practices for prompt engineering with OpenAI API", accessed: 22.05.2023, https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api.

[23] Dokumentacja Midjourney, accessed: 22.05.2023, https://docs.midjourney.com/docs/prompts.

SELEKTYWNE ZAKŁÓCANIE TRANSMISJI OFDM W INTERFEJSIE RADIOWYM NB-IOT

SELECTIVE JAMMING OF OFDM TRANSMISSION IN NB-IOT RADIO INTERFACE

Jarosław Magiera

Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, jaroslaw.magiera@pg.edu.pl

DOI: 10.15199/59.2023.4.78

Streszczenie**: W artykule przedstawiono zagadnienie tzw. inteligentnego zagłuszania, polegającego na selektywnym zakłócaniu krytycznych składowych sygnału OFDM. Opisane zostały badania weryfikujące możliwość przeprowadzenia tego rodzaju ataku elektronicznego. Przedstawiono zrealizowane stanowisko badawcze, umożliwiające zakłócanie rzeczywistych sygnałów NB-IoT w łączu w dół. Wyniki badań wykazują możliwość realizacji selektywnego zagłuszania i jego większą skuteczność w stosunku do zakłócania zaporowego.**

Abstract**: The article presents the concept of smart jamming, which involves selective interference aimed at critical components of the OFDM signal. Research verifying the feasibility of this type of electronic attack are described. A realized testbed for jamming real NB-IoT signals in a downlink is presented. The test results demonstrate the feasibility of realizing selective jamming and its greater effectiveness compared to barrage interference.**

Słowa kluczowe**: atak elektroniczny, inteligentne zagłuszanie, NB-IoT, OFDM.**

Keywords**: electronic attack, NB-IoT, OFDM, smart jamming.**

1. WSTĘP

Technika transmisji OFDM, ze względu na swoje liczne zalety, jak np. elastyczna alokacja zasobów czasowo-częstotliwościowych czy odporność na interferencje międzysymbolowe, znajduje szerokie zastosowanie we współczesnych interfejsach radiowych. W szczególności, OFDM jest fundamentem warstwy fizycznej sieci dostępowych RAN w standardach LTE i 5G.

Publikowane przez 3GPP specyfikacje techniczne warstw fizycznych interfejsów RAN precyzują organizację transmisji w blokach zasobów fizycznych (PRB - physical resource block) reprezentowanych przez określoną liczbę podnośnych i symboli OFDM. Dokumenty te definiują alokację zasobów w blokach dla potrzeb transmisji poszczególnych kanałów fizycznych. Część sygnałów i kanałów fizycznych ma statycznie przydzielone zasoby, tzn. ich transmisja odbywa się w ustalonych odstępach czasu w określonym zakresie podnośnych. Dotyczy to w szczególności transmisji rozgłoszeniowej w łączu w dół, umożliwiającej realizację procedury podłączenia się terminala użytkownika (UE – user equipment) do komórki. Taka organizacja transmisji zwiększa podatność interfejsu na celowe zakłócanie, gdyż uniemożliwienie łączności jest możliwe poprzez zagłuszenie jedynie wybranych elementów zasobów. Sprawia to, że urządzenie zagłuszające może działać bardziej efektywnie, niż gdyby emitowało nieprzerwanie zakłócenie obejmujące całe widmo sygnału użytecznego.

Do krytycznych elementów transmisji w łączu w dół zaliczają się: sygnały synchronizacyjne, sygnały referencyjne (pilotowe) oraz fizyczny kanał rozgłoszeniowy PBCH [1,2]. Zagłuszenie zasobów sygnałów synchronizacyjnych lub pilotowych uniemożliwi dostrojenie częstotliwości i taktu zegara po stronie UE lub, odpowiednio, uniemożliwi korekcję kanału. Z kolei pozbawienie możliwości odbioru kanału PBCH uniemożliwi pozyskanie przez UE informacji wymaganych do jego przyłączenia do komórki.

2. ZAKŁÓCANIE SELEKTYWNE

Urządzenia emitujące zakłócenia selektywne, w oparciu o znajomość struktury transmisji sygnału, są nazywane inteligentnymi zagłuszaczami (ang. smart jammer) [3]. Zakładając, że emisja zakłócenia odbywa się z ustaloną mocą szczytową, przy zwykłym zakłócaniu zaporowym, moc ta jest rozproszona w całym paśmie sygnału, natomiast w wariancie selektywnym może być skupiona na wybranych podnośnych. Skutkuje to większą skutecznością lub zasięgiem zakłócania. Ponadto, impulsowy charakter zakłóceń selektywnych zmniejsza zużycie energii potrzebnej do zasilania nadajnika.

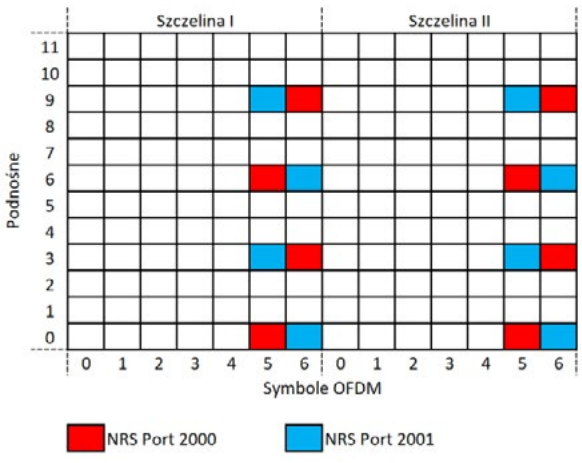
Realizacja zagłuszania selektywnego wymaga zapewnienia synchronizacji czasowej zakłócenia z taktem ramkowym sygnału użytecznego. W ogólności jest to możliwe poprzez wyposażenie stacji zakłócającej w odbiornik tego sygnału, z wyprowadzonym wyjściem zegara. Niemniej, takie podejście jest problematyczne z kilku względów. Przede wszystkim, emisja zakłócenia uniemożliwia równoczesny odbiór sygnału, co jest wymagane do podtrzymania synchronizacji. Po drugie, stacja zakłócająca musi znajdować się w zasięgu odbioru sygnału użytecznego. Wyklucza to scenariusz, w którym źródło zakłóceń, wyposażone w silnie kierunkową antenę, znajduje się w dużej odległości od zakłócanego odbiornika. Jakkolwiek, możliwe jest inne podejście, zakładające że zegary stacji bazowych sieci komórkowej są synchronizowane w oparciu o sygnały satelitarnych systemów nawigacyjnych GNSS. W takiej sytuacji, stacja zakłócająca może być również synchronizowana względem tych systemów i, dzięki temu, jej działanie jest niezależne od odbioru zakłócanego sygnału.

3. STRUKTURA TRANSMISJI NB-IOT

Badania opisane w dalszej części artykułu dotyczą zakłócania interfejsu radiowego NB-IoT, którego podatność na zagłuszanie była analizowana m.in. w [4,5]. NB-IoT jest jednym ze standardów zdefiniowanych przez 3GPP, dla potrzeb zapewnienia efektywnej wąskopasmowej komunikacji pomiędzy urządzeniami Internetu Rzeczy. Interfejs radiowy pomiędzy terminalem NB-IoT a RAN implementuje uproszczoną wersję stosu protokołów znanego z sieci LTE. Same zasoby NB-IoT mogą być wydzieloną częścią pasma LTE (tryb in-band) lub funkcjonować niezależnie (tryb standalone). Niezależnie od trybu, w łączu w dół stosowana jest transmisja OFDM obejmująca pojedynczy PRB o szerokości pasma 180 kHz (12 podnośnych po 15 kHz). Transmisja w łączu w dół zorganizowana jest w ramki o długości 10 ms, jedna ramka składa się z dziesięciu podramek, a każda podramka jest podzielona na dwie szczeliny przenoszące po 7 symboli OFDM. W strukturze transmisyjnej można wyróżnić charakterystyczne statyczne alokacje zasobów [6]:

- główny sygnał synchronizacyjny (NPSS) zajmuje zasoby podramki 5 w każdej ramce,
- dodatkowy sygnał synchronizacyjny (NSSS) jest obecny w podramce 9 każdej parzystej ramki,
- rozgłoszeniowy kanał fizyczny (NPBCH) jest przesyłany w podramce 0 każdej ramki,
- sygnał referencyjny NRS obejmuje dwie podnośne symbolu 5 oraz inne dwie podnośne symbolu 6 w każdej szczelinie, z wyjątkiem podramek zawierających sygnały NPSS i NSSS – przykładowe rozmieszczenie zasobów NRS w podramce przedstawiono na Rys. 1.

Pozostałe podramki są przeznaczone do przenoszenia sterującego kanału fizycznego NPDCCH oraz współdzielonego kanału fizycznego NPDSCH. W tym wypadku alokacja zasobów jest realizowana w sposób dynamiczny, zależnie od konfiguracji komórki oraz natężenia ruchu. Można, w związku z powyższym, stwierdzić, że najbardziej narażonymi na zakłócanie selektywne są podramki 0 i 5 oraz zasoby sygnału NRS.



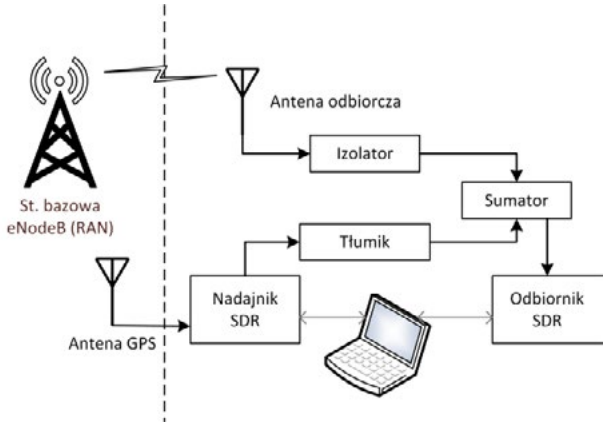
Rys. 1. Przykładowa alokacja zasobów OFDM dla sygnału NRS w jednej podramce.

4. BADANIA EMPIRYCZNE

Podjęte badania służą wykazaniu możliwości realizacji selektywnego zakłócenia sygnału OFDM, na przykładzie transmisji NB-IoT w łączu w dół. Ponadto, celem jest porównanie skuteczności zakłócenia selektywnego z zaporowym o jednakowej mocy szczytowej.

4.1. Stanowisko badawcze

W celu przeprowadzenia tych badań przygotowano stanowisko pomiarowe, którego schemat przedstawia Rys. 2.



Rys. 2. Schemat stanowiska badawczego.

Przebieg zakłócenia ma postać siatki OFDM, w której pozycje docelowo zakłócanych zasobów czasowo-częstotliwościowych są zajęte przez impulsy o losowej fazie i jednostkowej amplitudzie. W lokalizacjach pozostałych zasobów impulsy są nieobecne. Przebieg jest generowany za pomocą skryptu języka Matlab i zapisywany do pliku binarnego. Następnie plik ten jest przesyłany w pętli do radia programowalnego NI USRP X310, którego tor nadawczy wytwarza sygnał w.cz. na zadanej częstotliwości nośnej. Transmisja jest synchronizowana sygnałem PPS z wbudowanego w USRP odbiornika GPS.

Sygnałem zakłócanym jest transmisja NB-IoT z rzeczywistej stacji bazowej eNodeB. Wybór takiego źródła sygnału pozwala stwierdzić, czy założenie o synchronizacji zegara stacji bazowej względem zegara GPS jest słuszne. Nie byłoby to możliwe przy zastosowaniu testera radiokomunikacyjnego jako źródła sygnału NB-IoT.

Sygnał odbierany jest sumowany z zakłóceniem w pasmie w.cz. Pomiędzy nadajnikiem a sumatorem umieszczono tłumik regulowany, umożliwiający uzyskanie odpowiedniego stosunku mocy zakłócenia do mocy sygnału. Z kolei pomiędzy anteną odbiorczą a sumatorem zastosowano izolator zapobiegający emisji zakłócenia poza stanowisko badawcze.

Sygnał z wyjścia sumatora jest podawany na wejście toru odbiorczego radia programowalnego, w którym następuje zapis próbek do pliku. Dalsze przetwarzanie odbywa się z użyciem opracowanych skryptów Matlab do analizy i dekodowania sygnałów.

4.2. Scenariusz badań

Źródłem sygnału użytecznego dla potrzeb badań była stacja bazowa eNodeB sieci Plus, emitująca sygnał NB-IoT w trybie standalone, na częstotliwości

935.1 MHz, zlokalizowana na dachu jednego z budynków Politechniki Gdańskiej.

Pierwszym etapem badań była weryfikacja, czy transmisja sygnału ze stacji bazowej jest zsynchronizowana z zegarem systemu GPS. W szczególności, należało ustalić, czy z początkiem każdej sekundy zegara GPS rozpoczyna się transmisja ramki parzystej, tj. zawierającej sygnał NSSS w podramce 9. Aby to sprawdzić, wygenerowano testowy sygnał o charakterystycznym wzorcu czasowym i zarejestrowano go równocześnie z sygnałem użytecznym. Następnie porównano spektrogramy obu przebiegów.

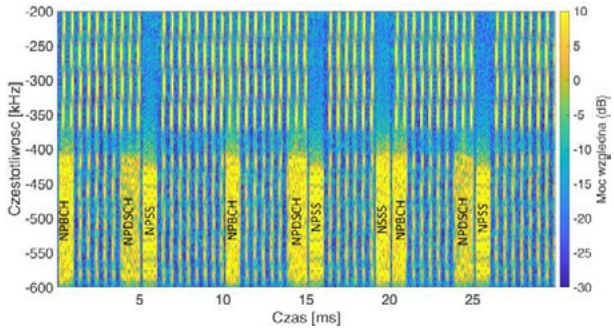
W drugim etapie badań porównano efektywność zakłócania selektywnego z zakłóceniem zaporowym, tj. ciągłym i obejmującym całe pasmo sygnału. Skupiono się na zakłócaniu kanału fizycznego NPBCH, przenoszącego główny blok informacyjny MIB (Master Information Block), czyli pierwsze informacje jakie odbiera UE podczas podłączenia do komórki. Zakłócenie selektywne miało postać impulsów zakłócających zasoby sygnału NRS w podramce 0 oraz poprzedzające je zasoby NRS (tj. zasoby NRS w drugiej szczelinie podramki 8 lub 9, w zależności od parzystości ramki). Z kolei zakłócenie zaporowe obejmowało całe pasmo sygnału NB-IoT we wszystkich podramkach, z wyłączeniem podramek sygnałów NPSS i NSSS (aby ograniczyć wpływ zakłócenia na zdolność synchronizacji UE).

4.3. Wyniki

W etapie 1 wygenerowano przebieg OFDM, w którym emisja jest aktywna wyłącznie w symbolach i podnośnych odpowiadających zasobom przeznaczonym dla sygnału NRS w sygnale użytecznym (w celu określenia lokalizacji tych zasobów, uprzednio ustalono identyfikator NB cell ID komórki). Następnie, na podstawie tego przebiegu, wygenerowano sygnał na częstotliwości 935.3 MHz, a więc przesuniętej o 200 kHz w górę względem sygnału użytecznego.

Na rys. 3. Przedstawiono spektrogram obejmujący pasmo częstotliwości o szerokości 400 kHz. W dolnej połowie tego pasma, na częstotliwości podstawowej -500 kHz, widoczne jest widmo sygnału odbieranego ze stacji bazowej. Oś czasu spektrogramu została ustalona w taki sposób, aby rozpoczynała się z początkiem ramki sygnału i obejmowała trzy kolejne ramki, tj. 30 ms. W widmie sygnału doskonale widoczne są charakterystyczne elementy skojarzone z poszczególnymi sygnałami i kanałami fizycznymi. W pierwszej, 11-tej i 21-szej milisekundzie występuje transmisja NPBCH. Z kolei w 6-tej, 16-tej i 26-tej – transmisja sygnału NPSS, a w 20-tej NSSS – wynika stąd, że pierwsza ramka w spektrogramie jest nieparzysta. Dodatkowo, w 5-tej, 15-tej i 25-tej milisekundzie widoczne są zajęte zasoby kanału NPDSCH, skojarzone z transmisją systemowego bloku informacyjnego (SIB). Pojedyncze impulsy widoczne na czterech podnośnych w pozostałych przedziałach czasu to elementy sygnału referencyjnego NRS.

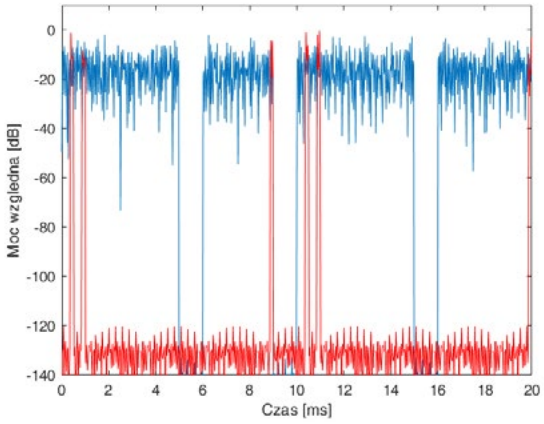
W górnej części spektrogramu, wokół częstotliwości podstawowej -300 kHz, jest natomiast widoczne widmo sygnału wygenerowanego. Jak można zauważyć, przedziały czasu, w których występują impulsy w sygnale testowym, pokrywają się z okresami transmisji symboli referencyjnych NRS w sygnale ze stacji bazowej. Ponadto, dłuższe przerwy w emisji wygenerowanego sygnału występują dokładnie w okresach nadawania sygnałów synchronizacyjnych NPSS i NSSS, co jest zgodne z oczekiwaniami. Opisane obserwacje dowodzą założenia, że sygnał stacji bazowej NB-IoT jest zsynchronizowany z zegarem systemu GPS.



Rys. 3. Spektrogram weryfikujący synchronizację sygnałów.

W etapie drugim badano efektywność zakłócania kanału NPBCH, zdefiniowaną jako odsetek niezdekodowanych bloków informacyjnych MIB, tzn. takich, w których wystąpiła niezgodność sumy kontrolnej CRC.

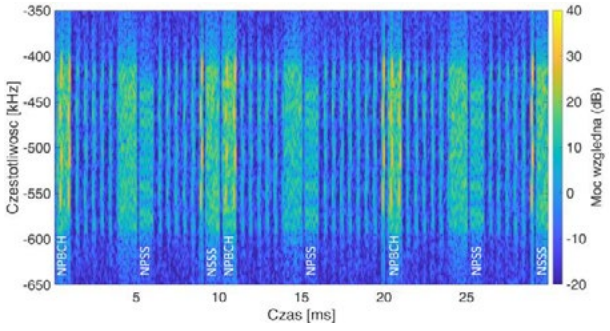
Na rys. 4. przedstawiono przebiegi mocy chwilowej zakłócenia zaporowego (niebieski) oraz selektywnego NRS (czerwony). Jak można zauważyć, moce szczytowe obu zakłóceń są na zbliżonym poziomie.



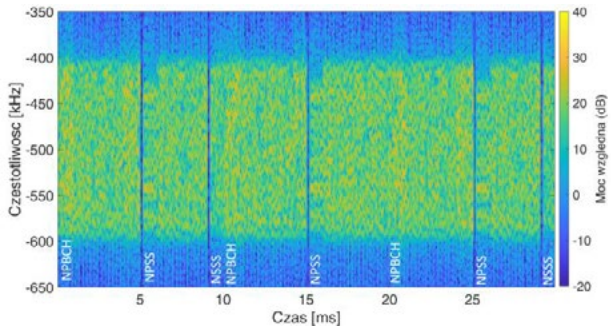
Rys. 4. Przebiegi mocy chwilowej zakłóceń.

Z kolei na rys. 5. i 6. uwidoczniono spektrogramy sygnałów zarejestrowanych w obecności zakłóceń, przy stosunku mocy szczytowej zakłócenia do mocy szczytowej sygnału wynoszącym ok. 3 dB. Dla ułatwienia interpretacji, oznaczono lokalizacje charakterystycznych podramek. W wypadku zakłócenia selektywnego, widoczne są jego impulsy występujące w miejscach symboli NRS w podramce NPBCH oraz przed nią. Z kolei w wypadku zakłócenia zaporowego, obejmuje ono cały sygnał z wyjątkiem podramek NPSS i NSSS rozpoczynających się charakterystyczną przerwą w transmisji w czasie trzech pierwszych symboli podramki.

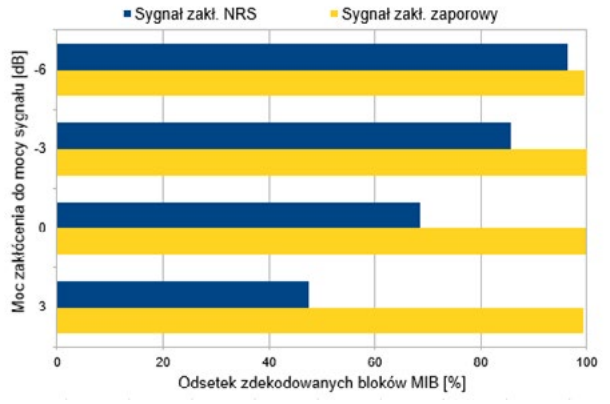
Na rys. 7. zamieszczono wyniki badania efektywności zakłócania kanału NPBCH. Przeprowadzono rejestrację sygnałów w obecności zakłóceń o stosunkach mocy szczytowej wynoszących -6 dB, -3 dB, 0 dB i 3 dB względem mocy szczytowej sygnału. W każdym przypadku podjęto próby zdekodowania 500 kolejnych bloków informacyjnych MIB. Jak można zauważyć, zakłócenie zaporowe o takiej mocy właściwie nie wpływa na zdolność dekodowania informacji MIB. Występowały jedynie pojedyncze przypadki niezdekodowania MIB, co jest możliwe nawet w wypadku braku zakłócenia. Odmiennie prezentują się wyniki dla zakłócenia selektywnego. Widać tu wyraźną zależność efektywności zakłócania od mocy. W wypadku, gdy moc szczytowa zakłócenia była dwukrotnie większa od mocy sygnału, zdekodowano mniej niż połowę bloków MIB.



Rys. 5. Spektrogram zakłócania selektywnego NRS.



Rys. 6. Spektrogram zakłócania zaporowego.



Rys. 7. Efektywność zakłócania kanału NPBCH.

5. PODSUMOWANIE

Przeprowadzone badania wykazały, że jest możliwa realizacja selektywnego zakłócania sygnałów OFDM emitowanych w interfejsie radiowym NB-IoT w łączu w dół. Stosowanie synchronizacji zegarów stacji bazowych sygnałami GNSS umożliwia wykorzystanie tychże sygnałów do wyzwalania emisji w urządzeniu zakłócającym, bez potrzeby uprzedniego odbioru przez nie sygnału użytecznego.

Zdecydowanie większa efektywność zagłuszania selektywnego wynika głównie ze skupienia mocy zakłócenia na wybranych podnośnych. W analizowanym przypadku, zakłócenie było obecne równocześnie jedynie na dwóch podnośnych NRS, podczas gdy zakłócenie zaporowe obejmowało aż dwanaście podnośnych. Zakłócenie skupione na zasobach NRS powodowało zaburzenia w działaniu procedury estymacji współczynników kanału dla podramki 0, co w konsekwencji uniemożliwiało zdekodowanie bloku MIB.

LITERATURA

[1] Y. Arjoune and S. Faruque, *Smart Jamming Attacks in 5G New Radio: A Review*, 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2020, pp. 1010-1015, doi: 10.1109/CCWC47524.2020.9031175.

[2] Skokowski, P., Kelner, J. M., Malon, K., Maślanka, K., Birutis, A., Vazquez, M. A., ... Ambroziak, S. (2022*). Jamming and jamming mitigation for selected 5G military scenarios*. Procedia Computer Science, 205, 258-267.

[3] Topal, O. A., Gecgel, S., Eksioglu, E. M., & Kurt, G. K. (2019). *Identification of smart jammers: Learning based approaches using wavelet representation*. arXiv preprint arXiv:1901.09424..

[4] Morillo, G., & Roedig, U. (2021, October). *Jamming of NB-IoT synchronisation signals.* 26th European Symposium on Research in Computer Security (ESORICS) 2021 (pp. 1-5).

[5] Staniec, K., & Kowal, M. (2020*). On vulnerability of selected IoT systems to radio jamming—A proposal of deployment practices*. Sensors, 20(21), 6152.

[6] Specyfikacja techniczna 3GPP TS 36.211, *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation*, Release 14

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

Zastosowanie Heurystyki do Wyboru Lokalizacji Węzłów Pośredniczących w Sieciach LP WAN

Heuristic Approach for Relay Location Selection in LP WAN

Anna Strzoda[1]

[1] IITiS PAN, Gliwice, astrzoda@iitis.pl

DOI: 10.15199/59.2023.4.79

Streszczenie: Przedmiotem pracy jest propozycja algorytmu wyboru węzłów pośredniczących w sieciach LP WAN dla urządzeń końcowych znajdujących się w położeniu o niekorzystnych warunkach transmisyjnych powodujących utratę pakietów. W zaproponowanym podejściu zastosowano metodę heurystyczną mającą na celu znalezienie rozwiązania zoptymnalizowanego pod względem oszczędzania energii urządzeń sieci.

Abstract: The subject of the study is the proposal of an algorithm for selecting relay nodes in LP WAN networks for end devices located in positions with inconvenient transmission conditions causing packet loss. The proposed approach utilizes a heuristic method to find an energy-efficient optimized solution.

Słowa kluczowe: 2-hop, LoRa, LoRaWAN, LP WAN, Relay, Węzły pośredniczące

Keywords: 2-hop, LoRa, LoRaWAN, LP WAN, Relay

1. WSTĘP

Low Power Wide Area Network (LPWAN) to zestaw technologii komunikacji bezprzewodowej stworzonych w celu umożliwienia przesyłania niewielkiej ilości danych na duże odległości przy niskim poborze energii. Standard LoRaWAN jest jedną z technologii LPWAN działających w paśmie nielicencjonowanym (np. 868 MHz w Unii Europejskiej, 902-928 MHz w USA). LoRaWAN korzysta z modulacji rozproszonego widma CSS (chirp spread spectrum). LoRaWAN wykorzystuje różne współczynniki rozprzestrzeniania (ang. spreading factor): SF7, SF8, SF9, SF10, SF11, SF12, które determinują liczbę symboli wysłanych w czasie jednej sekundy. Im niższa wartość SF, tym szybsza transmisja danych oraz mniejsze zużycie energii, ale także mniejszy zasięg.

Technologia ta znajduje szerokie zastosowanie w IoT zwłaszcza w rozwiązaniach, które wymagają okresowego monitorowania za pomocą różnych czujników. Zestaw urządzeń sieciowych składa się z stacji bazowych LoRa (bram docelowych) oraz urządzeń innych niż stacje bazowe (źródła), przy czym te ostatnie z kolei podzielone są na kilka grup. Zbiór urządzeń niebędących stacją bazową obejmuje grupę urządzeń, które posiadają słabą łączność ze stacją bazową, co skutkuje utratą części pakietów. Z tego powodu istnienie urządzeń o słabej łączności warunkuje istnienie innej grupy urządzeń, jaką są urządzenia pośredniczące, których rolą jest retransmisja sygnałów ze źródła o słabej łączności do bramy docelowej.

Większość sieci LP WAN opiera się na topologii gwiazdy, gdzie bezpośrednia komunikacja zachodzi pomiędzy urządzeniami końcowymi a bramami LoRa. Urządzenia końcowe w sieci LoRa mogą mieć ograniczony zasięg, co może prowadzić do utraty połączenia między węzłami a bramami. Szczególnie w obszarach o gęstej zabudowie lub trudnym terenie. W związku z powyższym, rola węzłów pośredniczących w sieciach LP WAN jest tematem dyskutowanym w literaturze naukowej [19, 8] oraz w praktyce - pod koniec 2022 roku stowarzyszenie LoRa Alliance opublikowało rozszerzenie warstwy łącza LoRaWAN o specyfikację trybu relay [1].

Wybór węzłów pośredniczących oraz ich lokalizacji jest istotnym zagadnieniem mającym wpływ na wydajne działanie sieci. Rolę węzłów pośredniczących mogą pełnić zwykłe węzły końcowe sieci poprzez przekazywanie dalej pakietów z urządzeń, które w przeciwnym razie nie byłyby w stanie komunikować się z siecią. Takie podejście eliminuje potrzebę instalacji dodatkowych urządzeń i ułatwia przekazywanie danych przez sąsiednie urządzenia, które mają stabilne połączenie z bramą LoRa.

W niniejszej pracy analizowana jest kwestia wyboru węzłów pośredniczących dla urządzeń, które nie są w stanie bezpośrednio komunikować się z bramami LoRa oraz identyfikowany jest obszar działania urządzeń wskazanych do pracy w trybie relay. Opisano tutaj algorytm oparty o metodę heurystyczną dla znalezienia optymalnego przypisania węzła pośredniczącego dla każdego węzła, który znajduje się poza zasięgiem bezpośredniej komunikacji z bramami. Ewaluację metody przeprowadzono z wykorzystaniem środowiska symulacyjnego oraz w odniesieniu do metody referencyjnej dostępnej w literaturze.

2. PRZEGLĄD LITERATURY

Znacząca liczba prac badawczych wskazuje na zastosowanie przekazywania danych przez węzły pośredniczące w sieciach LP WAN. W pracy [18] pokazano, że możliwe jest zastosowanie węzła pośredniczącego LoRa zasilanego bateryjnie. W [3] autorzy proponują schemat komunikacji z węzłem pośredniczącym w celu poprawy niezawodności sieci czujników działających w technologii LoRa z uwzględnieniem tzw. limitu duty-cycle. Wykorzystanie węzłów pośredniczących zbadano również w innych sieciach LP WAN - na przykład opisano wykorzystanie przekaźników montowanych w pojazdach w sieci NB IoT [16]. Także przekazywanie danych z LP WAN do komunikacji satelitarnej jest opisane w [15]. Autorzy [5] wykazali,

że wybór węzłów pośredniczących w sieciach bezprzewodowych może być realizowany przy niewielkiej złożoności obliczeniowej i obciążenia systemów.

Problem wyboru węzłów pośredniczących został także zbadany, np. w [13], gdzie rozważa się wykorzystanie potencjału obliczeniowego urządzeń fogowych do pełnienia roli węzłów pośredniczących w ogólnym ekosystemie IoT. Autorzy proponują złożony, wielopoziomowy, iteracyjny algorytm minimalizujący prawdopodobieństwo awarii urządzenia. W [4] opisano podejście do projektowania wytrzymałych, bezprzewodowych sieci przekaźnikowych opartych na warstwie fizycznej LoRa do przesyłania krótkich i rzadkich komunikatów, takich jak dane z aktuatorów. W [12] zaproponowano mechanizm, który ma na celu wybór węzła pośredniczącego w geograficznym przekazywaniu pakietów w dużej sieci bezprzewodowej (WSN) z cyklem uśpienia i wybudzenia. Nie można go jednak bezpośrednio zastosować w sieciach LP WAN ze względu na konieczną wymianę sond, co prowadziłoby do dużej zajętości pasma ze względu na wysoki zasięg transmisji. W pracy [11] rozważono wybór węzłów pośredniczących nastawiony na oszczędzanie energii. Proponowany algorytm, PRS-D2D, wykorzystuje metodę węgierską do rozwiązania problemu przyporządkowania węzłów pośredniczących w czasie wielomianowym. Jednakże, rozwiązanie jest dedykowane dla wykorzystania z antenami kierunkowymi.

Większość prac nie rozważa zagadnienia wyboru lokalizacji węzłów pośredniczących w sieciach LP WAN. Najczęściej wskazanie urządzenia do pełnienia roli węzła pośredniczącego nie jest rezultatem procedury optymalizacyjnej uwzględniającej spektrum czynników takich jak położenie geograficzne czy oszczędzanie energii. Nie mniej jednak, w pracy [7] zaprezentowano kompleksowy algorytm wyboru węzłów pośredniczących dedykowany dla standardu LoRa, który wykazuje lepsze rezultaty pod kątem oszczędzania energii w sieci niż metoda [10], co pokazano z wykorzystaniem środowiska symulacyjnego oraz rzeczywistych topologii sieci. Niektóre z wcześniejszych prac badawczych analizują wybór węzłów pośredniczących także w siach komórkowych [9]. Autorzy [10] proponują aukcyjną optymalizację wyboru węzłów pośredniczących w tej technologii. Jest to metoda podobna do podejścia zaprezentowanego w niniejszej pracy ze względu na założenie, że jeden węzeł pośredniczący może zostać przyporządkowany do maksymalnie jednego urządzenia posiadającego słabą łączność z bramą docelową.

3. ALGORYTM

Dla znalezienia zestawu węzłów pośredniczących, tak aby zapewnić pełne pokrycie wszystkich urządzeń końcowych, które mają słabą łączność z bramą docelową LoRa i jednocześnie zapewnić optymalne zużycie energii w sieci, zastosowano metodę heurystyczną opartą o optymalizację kolonii mrówek (Ant Colony Optimization - ACO) [6, 2].

Topologię sieci reprezentuje nieskierowany graf ważony $G = (V_G, E_G, \Omega_G)$, ze zbiorem wierzchołków V_G, zbiorem krawędzi E_G i zbiorem wag krawędzi Ω_G. W niniejszym opisie każdy wierzchołek reprezentuje urządzenie sieciowe. Pomiędzy wierzchołkami występuje krawędź, jeśli odpowiadające im urządzenia znajdują się w swoim zasięgu. Każdej krawędzi $(v, w) \in E_G$ przypisana jest waga $SF_{vw} \in \{7, 8, 9, 10, 11, 12\} \subset \Omega_G$ reprezentująca wartość współczynnika rozprzestrzeniania w komunikacji między urządzeniami odpowiadającymi węzłom v i w.

Problem wybru zbioru węzłów pośredniczących można sprowadzić do problemu z teorii grafów, który polega na znalezieniu maksymalnego ważonego dopasowania w grafie dwudzielnym [14, 17]. Maksymalnie ważone dopasowanie w grafie dwudzielnym to problem znalezienia w ważonym grafie dwudzielnym dopasowania, w którym suma wag jest maksymalna.

W rozważanej topologii sieci wyszczególnia się zbiór potencjalnych węzłów pośredniczących, czyli takich, które posiadają nadwyżkę energii, nie pełnią roli węzła pośredniczącego dla żadnego węzła słabego oraz mają w zbiorze wierzchołków reprezentujących niepokryte węzły słabe co najmniej jednego sąsiada. Niepokryty węzeł słaby to taki, któremu należy przyporządkować węzeł pośredniczący.

Rozważa się ważony graf dwudzielny $H = (U, P, E_H, \Omega_H)$, gdzie zbiór U zawiera niepokryte węzły słabe, P jest zbiorem wierzchołków mających potencjał do pełnienia roli węzłów pośredniczących, $E_H \subset E_G$ to zbiór krawędzi pomiędzy węzłami słabymi a potencjalnymi węzłami pośredniczącymi, gdzie każdej krawędzi $(v, w) \in E_H$ jest przypisana waga $\eta(v, w) \in \Omega_H$, która została wyzaczona według formuły 4.

Celem algorytmu wyboru węzłów pośredniczących jest znalezienie przyporządkowania w ważonym grafie dwudzielnym H, takiego że każdemu węzłowi słabemu zostanie przyporządkowany dokładnie jeden węzeł pośredniczący, przy założeniu, że każdy węzeł pośredniczący może zostać przyporządkowany tylko jeden raz. Znalezione przyporządkowanie jest zmaksymalizowane pod kątem sumy wag krawędzi.

Stąd problem wyboru węzłów pośredniczących można sformułować w następujący sposób:

$$\max \sum_{(v,w)\in E_H} \eta(v, w) x_{vw} \tag{1}$$

$$\sum_{w\in U} x_{vw} \leqslant 1, \forall v \in P \tag{2}$$

$$\sum_{v\in P} x_{vw} \leqslant 1, \forall w \in U \tag{3}$$

gdzie $x_{vw} \in \{0, 1\}$ i x_{vw} równe 1 oznacza, że krawędź (v, w) jest krawędzią pomiedzy węzłem słabym i węzłem pośredniczącym w znalezionym przyporządkowaniu. Jeśli $|U| = |P|$, wówczas $\sum_{w\in U} x_{vw} = 1, \forall v \in P$ i $\sum_{v\in P} x_{vw} = 1, \forall w \in U$.

3.1 Funkcja heurystyczna

Funkcja heurystyczna służąca do oceny atrakcyjności wyboru węzła v jako węzła pośredniczącego dla węzła słabego w dana jest następującym wzorem:

$$\eta(v, w) = \alpha \times \gamma_{vw} + (1 - \alpha) \times \sigma_v \tag{4}$$

gdzie $\gamma_{vw} \in [0, 1]$ jest wynikiem normalizacji wartości parametru t_{vw} według następującej formuły:

$$\gamma_{vw} = \frac{t_{vw} - a}{b - a} \times (b_{new} - a_{new}) + a_{new}, \tag{5}$$

gdzie $a = \min_{(v,w)\in E_H} t_{vw}$, $b = \max_{(v,w)\in E_H} t_{vw}$, $a_{new} = 0$, $b_{new} = 1$. Wartość parametru t_{vw} została wyznaczona według rachunku energetycznego przedstawionego w pracy [7] i jest liczbą dni przez które węzeł v może pełnić funckję węzła pośredniczącego dla węzła w, tak aby

nie doprowadzić do wyczerpania baterii węzła pośredniczącego. Rachunek energetyczny z [7] uwzględnia wartość współczynnika rozprzestrzeniania, bieżący poziom baterii urządzenia, a także czas operacyjny urządzenia i stanowi podstawę algorytmu wyboru węzła pośredniczącego zaproponowanego przez autorów. Także wartość współczynnika rozprzestrzeniania v_{SF_g} w komunikacji pomiędzy potencajlnym węzłem pośredniczącym v i stacją bazową g przedstawia się jako $\sigma_v \in [0, 1]$ i wyznacza według formuły:

$$\sigma_v = 0.2 \times (12 - v_{SF_g}) \tag{6}$$

Funkcja heurystyczna (4) ocenia atrakcyjność połączenia między słabym węzłem w i potencjalnym węzłem pośredniczącym v. Funkcja ta składa się z dwóch głównych komponentów, biorąc pod uwagę długość czasu, przez jaki węzeł v może pełnić rolę węzła pośredniczącego dla słabego węzła w (γ_{vw}) oraz wartość współczynnika rozprzestrzeniania w komunikacji pomiędzy potencjalnym węzłem pośredniczącym v a docelową bramą LoRa (σ_v).

Rolą komponentu γ_{vw} jest wyrażenie atrakcyjności węzła v jako urządzenia pośredniczącego dla słabego węzła w pod względem czasu pracy w trybie węzła pośredniczącego do obsługi słabego węzła w. Im dłuższy czas pracy w trybie relay tym lepszy kandydat dla wskazanego słabego węzła do pełnienia roli węzła pośredniczącego. Do roli węzła pośredniczącego preferowane są węzły, które mogą pełnić tę funkcję dla poszczególnych słabych węzłów słąbych tak długo, jak to możliwe. Takie podejście zapewnia zmniejszenie liczby wyjść z trybu węzła pośredniczącego w przyszłości.

Drugi składnik σ_v wyraża ocenę węzła v na podstawie wartości współczynnika rozprzestrzeniania v_{SF_g} w komunikacji między węzłem v a bramą docelową g. Gdy wartość v_{SF_g} maleje, wartość σ_v rośnie. Takie podejście uwzględnia wybór węzła pośredniczącego ukierunkowany na oszczędzanie energii. Im mniejszy współczynnik rozprzestrzeniania się w komunikacji pomiędzy węzłem pośredniczącym a bramą docelową v_{SF_g}, tym mniej energii zużywa urządzenie reprezentowane przez węzeł v. Im wyższa wartość funkcji heurystycznej, tym bardziej atrakcyjny jest węzeł v do pełnienia roli węzła pośredniczącego dla słabego węzła w.

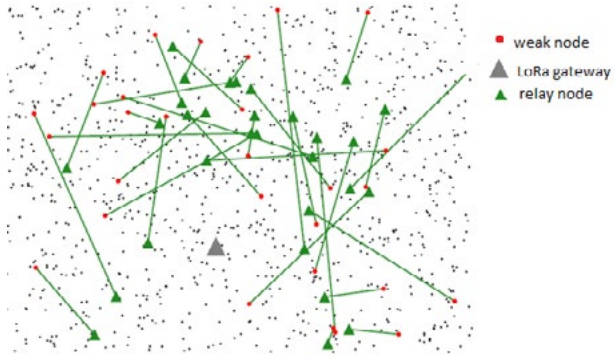
Wielkość współczynnika α determinuje wagę z jaką zostaną uwzględnione komponenty γ_{vw} oraz σ_v w procesie znajdowania optimum. W eksperymentach opisanych w rozdziale 4 przyjęto wielkość parametru $\alpha = 0.5$. Niniejsza wartość powoduje uwzględnienie obydwu komponentów tj. γ_{vw} oraz σ_v jako jednakowo ważnych.

4. EWALUACJA METODY

Do ewaluacji zaproponowanej metody przeprowadzono test na symulacyjnym stanowisku testowym w środowisku OMNeT++ zaczerpniętym wprost z [7], gdzie znajduje się jego szczegółowy opis oraz wartości parametrów symulacji. Autorzy pracy wykorzystali ów model do ewaluacji ich własnej metody wyboru węzłów pośredniczących.

Uwzględniono dwa scenariusze losowego rozmieszczenia węzłów w przestrzeni

- Scenariusz A: liczba węzłów sieci: 1000, obszar: 1000m × 1500m, odsetek węzłów słabych: 3%
- Scenariusz B: liczba węzłów sieci: 1500, obszar: 2000m × 2500m, odsetek węzłów słabych: 3%



Rysunek 1: Wizualizacja przyporządkowania węzłów pośredniczących w losowej topologii sieci. Liczba węzłów: 1000, obszar 1000m × 1500m, odsetek węzłów słabych: 3%.

Dla każdego scenariusza z osobna uruchomiono - niezależnie - algorytm mrówkowy oraz metodę referencyjną [10] wyboru węzłów pośredniczących, uzyskując tym samym dwa rozwiązania w postaci wyniokowych zbiorów węzłów pośredniczących w obrębie pojedynczego scenariusza. Następnie, przeprowadzono dwie oddzielne symulacje działania sieci, które różniły się zestawem węzłów pośredniczących w zależności od zastosowanej metody: algorytmu mrówkowego oraz metody referencyjnej. Celem badań było wyznaczenie całkowitego zużycia energii w sieci dla analizy porównawczej metod wyboru węzłów pośredniczących pod kątem oszczędzania energii. W symulacjach uwzględniono okres 10 lat działania sieci i dokonano pomiaru zużycia energii przez wszystkie urządzenia w tym czasie.

Wyniki eksperymentów pokazują, że różnica w zużyciu energii sieci dla zaproponowanego algorytmu wyboru węzła pośredniczącego oraz metody referencyjnej jest nieznacząca. W obydwu scenariuszach większą oszczędność energii zaobserwowano dla sieci z zestawem węzłów pośredniczących wskazanych przez algorytm referencyjny [10]: o 0.0007% dla scenariusza A oraz o 0.0024% dla scenariusza B. Obydwa podejścia są podonbne do siebie pod względem rodzaju przyporządkowania tzn. że każdemu węzłowi słabemu przypisany zostaje dokładnie jeden węzeł pośredniczący, a każdy węzeł pośredniczący pełni rolę przekaźnika dla dokładnie jednego węzła słabego.

Czas działania zaproponowanej metody optymalizacyjnej wyboru węzłów pośredniczących zależy od rozmiaru danych wejściowych oraz liczby iteracji wykonanych w procesie szukania rozwiązania. Algorytm przerywa proces przeszukiwania przestrzeni rozwiązań w momencie znalezienia rozwiązania optymalnego, tzn. którego wartość oceny jest równa maksymalnej wartości oceny możliwej do osiągnięcia. W przyjętych scenariuszach czas działania zaproponowanej metody nie przekracza 45 sekund.

5. KONKLUZJE

Przedstawiono heurystyczną metodę wyboru węzłów pośredniczących, dedykowaną dla standardu LoRa. Zaprezentowane podejście uwzględnia położenie geograficzne urządzeń sieci i ma na celu optymalizację pod względem zużycia energii urządzeń sieci.

Otrzymane rezultaty wykazują, że skuteczność zaprezentowanego algorytmu jest porównywalna do metody dostępnej w literaturze [10] pod względem zużycia ener-

gii, co zostało potwierdzone w środowisku symulacyjnym. Obie metody charakteryzuje przyporządkowanie w którym każdy węzeł pośredniczący zostaje przypisany dokładnie jednemu węzłowi słabemu.

Nie mniej jednak, zaproponowane podejście jest rozwiązaniem dopasowanym do standardu LoRa co daje perpesktywę do pracy nad jego rozwojem w tej technologii, w przeciwieństwie do metody referencyjnej powstałej dla technologii sieci komórkowych.

LITERATURA

[1] LoRa Alliance. Lorawan relay specification ts011-1.0.0. https://resources.lora-alliance.org/technical-specifications/ts011-1-0-0-relay, 2022. [Online; accessed 17-May-2023].

[2] Christian Blum. Ant colony optimization: Introduction and recent trends. *Physics of Life reviews*, 2(4):353–373, 2005.

[3] Siddhartha S. Borkotoky, Udo Schilcher, and Christian Bettstetter. Cooperative relaying in lora sensor networks. In *2019 IEEE Global Communications Conference (GLOBECOM)*, pages 1–5, 2019.

[4] Philip Branch and Tony Cricenti. A lora based wireless relay network for actuator data. In *2020 International Conference on Information Networking (ICOIN)*, pages 190–195. IEEE, 2020.

[5] Jun Cai, Xuemin Shen, Jon W. Mark, and Attahiru S. Alfa. Semi-distributed user relaying algorithm for amplify-and-forward wireless relay networks. *IEEE Transactions on Wireless Communications*, 7(4):1348–1357, 2008.

[6] Marco Dorigo, Mauro Birattari, and Thomas Stutzle. Ant colony optimization. *IEEE Computational Intelligence Magazine*, 1(4):28–39, 2006.

[7] Krzysztof Grochla, Anna Strzoda, Rafał Marjasz, Przemysław Głomb, Kamil Książek, and Zbigniew Łaskarzewski. Energy-aware algorithm for assignment of relays in lp wan. 18(4), 2022.

[8] Qinwei He, Tian Lan, Jianqi Li, Xiaopeng Yuan, and Yulin Hu. A wireless relay assisted lpwan for condition monitoring of converter stations. In *2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP)*, pages 898–902. IEEE, 2021.

[9] Sachin Kadloor and Raviraj Adve. Relay selection and power allocation in cooperative cellular networks. *IEEE Transactions on Wireless Communications*, 9(5):1676–1685, 2010.

[10] Yun Li, Chao Liao, Yong Wang, and Chonggang Wang. Energy-efficient optimal relay selection in cooperative cellular networks based on double auction. *IEEE Transactions on Wireless Communications*, 14(8):4093–4104, 2015.

[11] B. Ma, H. Shah-Mansouri, and V. W. S. Wong. A matching approach for power efficient relay selection in full duplex d2d networks. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–6, 2016.

[12] K. P. Naveen and Anurag Kumar. Relay selection with channel probing in sleep-wake cycling wireless sensor networks. *ACM Trans. Sen. Netw.*, 11(3), may 2015.

[13] B. Omoniwa, R. Hussain, M. Adil, A. Shakeel, A. K. Tahir, Q. U. Hasan, and S. A. Malik. An optimal relay scheme for outage minimization in fog-based internet-of-things (iot) networks. *IEEE Internet of Things Journal*, 6(2):3044–3054, 2019.

[14] C.N.K. Osiakwan and S.G. Akl. The maximum weight perfect matching problem for complete weighted graphs is in pc. In *Proceedings of the Second IEEE Symposium on Parallel and Distributed Processing 1990*, pages 880–887, 1990.

[15] Maria Rita Palattella and Nicola Accettura. Enabling internet of everything everywhere: Lpwan with satellite backhaul. In *2018 Global Information Infrastructure and Networking Symposium (GIIS)*, pages 1–5. IEEE, 2018.

[16] Vitaly Petrov, Andrey Samuylov, Vyacheslav Begishev, Dmitri Moltchanov, Sergey Andreev, Konstantin Samouylov, and Yevgeni Koucheryavy. Vehicle-based relay assistance for opportunistic crowdsensing over narrowband iot (nb-iot). *IEEE Internet of Things journal*, 5(5):3710–3723, 2017.

[17] Olfa Sammoud, Christine Solnon, and Khaled Ghédira. Ant algorithm for the graph matching problem. In *Evolutionary Computation in Combinatorial Optimization: 5th European Conference, EvoCOP 2005, Lausanne, Switzerland, March 30-April 1, 2005. Proceedings 5*, pages 213–223. Springer, 2005.

[18] ALESSIO Sanfratello, Enzo Mingozzi, and Francesco Marcelloni. Enabling relay-based communication in lora networks for the internet of things: design implementation and experimental evaluation. *Pisa, Italy*, 2016.

[19] Wenyang Xu, Guofa Cai, Yi Fang, and Guanrong Chen. Performance analysis of a two-hop relaying lora system. In *2021 IEEE/CIC International Conference on Communications in China (ICCC)*, pages 540–545. IEEE, 2021.

INFRASTRUKTURA SYSTEMU DO MONITOROWANIA AKTYWNOŚCI OSÓB STARSZYCH

INFRASTRUCTURE OF THE ELDERLY ACTIVITY MONITORING SYSTEM

Wojciech Hrycenko[1]; Michał Twarowski[2]; Vitomir Djaja-Jośko[3]

[1] Politechnika Warszawska, Warszawa, Wojciech.Hrycenko.stud@pw.edu.pl
[2] Politechnika Warszawska, Warszawa, Michal.Twarowski3.stud@pw.edu.pl
[3] Politechnika Warszawska, Warszawa, Vitomir.Djaja-Josko@pw.edu.pl

DOI: 10.15199/59.2023.4.80

Streszczenie: **Według badań, wśród starszych osób coraz częściej obserwowany jest spadek motywacji do podejmowania działań i pogorszenie zdrowia. Potencjalnym sposobem na zwiększenia ich jakości życia jest rozwijanie systemów ICT, działających zgodnie z paradygmatami Positive Health i Healthy Aging. Taki system jest opracowywany w ramach projektu CAREUP. Jednym z jego zadań ma być monitorowanie aktywności osób starszych z wykorzystaniem zestawu czujników środowiskowych oraz urządzeń noszonych (etykiet). Centralną częścią systemu jest infrastruktura złożona z węzłów odbierających dane od czujników i etykiet oraz bramki przekazującej dane do chmury (lub serwera) z wykorzystaniem łącza LTE-M. W referacie przedstawiono architekturę systemu i opisano sterowniki węzłów oraz bramki gromadzącej i przetwarzającej dane.**

Abstract: **According to the recent research, an increasing decline in the motivation to take actions and health deterioration is observed among seniors. One of the possible ways of improving the seniors' quality of life is utilization of the ICT-based assisting-systems working in line with the Positive Health and Healthy Aging paradigms. Such system is developed within the CAREUP project. One of its main goals is activity monitoring of the seniors using the set of the environmental sensors and wearables (tags). Core part of the system is its infrastructure that consist of the anchor nodes responsible for the reception of the data sent by the sensors and tags and the gateway relaying the data to the cloud (or server) using the LTE-M link. In the paper system's architecture is presented together with the description of the anchors and the gateway.**

Słowa kluczowe: **BLE, LTE-M, monitorowanie aktywności**

Keywords: **activity monitoring, BLE, LTE-M.**

1. WSTĘP

Żyjemy w starzejącym się społeczeństwie. Według szacunków w Europie w 2060 roku osoby w wieku powyżej 65 lat będą stanowiły ok. 28% całej populacji [5]. Postępy, jakie czyni współczesna medycyna sprawiają, że ludzie żyją dłużej. Niestety nie zawsze wiąże się to z zachowaniem pełni zdrowia, czy całkowitej samodzielności. Wiele z tych osób może cierpieć na różnego rodzaju choroby i wymagać opieki. W większości przypadków (ok. 80%) taką opiekę zapewniają rodzina, przyjaciele bądź pracownicy sektora opieki zdrowotnej [15]. Widać więc, że konieczne jest opracowywanie systemów zapewniających wsparcie tym osobom. W literaturze można znaleźć różne przykłady takich systemów. Warto wśród nich wyróżnić system NITICS [2], będący platformą integrującą systemy automatyki domowej, system lokalizacyjny oraz interfejsy do powiadamiania opiekunów.

Niemniej jednak, nawet osobom starszym nietkniętym żadnymi dolegliwościami powinna być dostarczona opieka, w ramach której można wyróżnić dwa podstawowe cele:

- zapewnienie poczucia bezpieczeństwa poprzez monitorowanie aktywności i stanu zdrowia osoby objętej opieką, aby móc na bieżąco wykrywać sytuacje potencjalnie zagrażające życiu (np. upadek),
- długofalowe analizowanie danych dotyczących aktywności i danych medycznych (np. ciśnienia krwi, wagi, poziomu cukru) w celu wczesnego wykrywania ewentualnych zmian chorobowych.

W kwestii systemów przeznaczonych do monitorowania aktywności osób starszych można wyróżnić trzy podstawowe kategorie systemów:

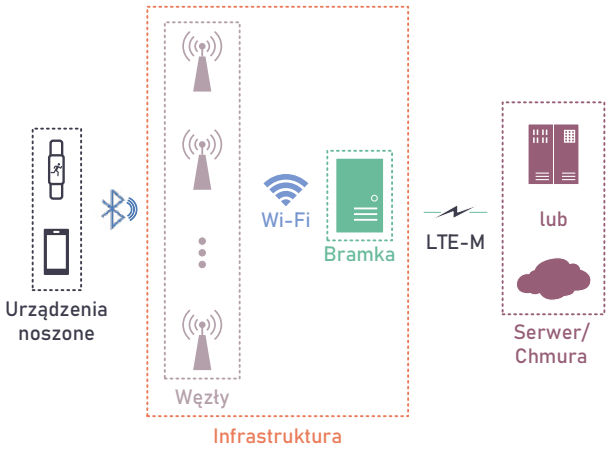
- wizyjne, których działanie oparte jest na kamerach [6], co może wzbudzać wątpliwości w kwestii prywatności oraz budzić pytania natury etycznej,
- oparte na czujnikach instalowanych w domach, przykładowo czujnikach zbliżeniowych czy podczerwieni lub elementach automatyki domowej (np. czujnikach otwierania drzwi) [3], co nie zawsze może zapewniać pożądaną dokładność,
- oparte na czujnikach noszonych przez osoby starsze [14], co zapewnia dużą dokładność, ale nie zawsze może być komfortowe, szczególnie przy długotrwałym użytkowaniu.

W niniejszym referacie skupiono się na koncepcji infrastruktury systemu do monitorowania aktywności starszych osób z wykorzystaniem urządzeń noszonych (np. inteligentnych opasek) wyposażonych w czujniki inercyjne.

Struktura referatu jest następująca. W rozdziale drugim przedstawiono ogólną architekturę omawianego systemu. W rozdziałach trzecim i czwartym opisano odpowiednio sterowniki węzłów oraz bramki. Natomiast referat podsumowano w rozdziale piątym.

2. OGÓLNA ARCHITEKTURA SYSTEMU

Na rys. 1 przedstawiono ogólną architekturę systemu do monitorowania aktywności osób starszych.



Rys. 1 Ogólna architektura systemu do monitorowania aktywności

System składa się z trzech podsystemów – urządzeń noszonych przez osoby starsze (np. inteligentnych opasek lub etykiet), infrastruktury, której zadaniem jest odbieranie danych od urządzeń noszonych (interfejs BLE), ich wstępne przetwarzanie oraz przekazywanie do chmury lub serwera (interfejs LTE-M) na których zaimplementowana jest cała logika sterująca pracą systemu, analizująca na bieżąco dane, podejmująca na ich podstawie pewne decyzje oraz dostarczająca informacje do ewentualnych interfejsów użytkownika.

Infrastruktura składa się z węzłów oraz bramki. Węzły powinny być wyposażone w interfejsy radiowe (BLE i Wi-Fi), umożliwiające komunikację zarówno z etykietami, jak i z bramką, natomiast bramka powinna być wyposażona w modem komórkowy (LTE-M) umożliwiający przesyłanie danych do serwera lub chmury.

Taka architektura umożliwia elastyczny, dostosowany do potrzeb danej instalacji systemu, dobór liczby węzłów, z zachowaniem jednej bramki agregującej pochodzące z nich dane (w tym celu wykorzystano interfejs Wi-Fi), dokonującej wstępnych obliczeń i analizy danych i przesyłającej je za pomocą łącza LTE-M do serwera bądź chmury (co uniezależnia instalację systemu od innych interfejsów komunikacyjnych dostępnych w domu osoby starszej). Z kolei interfejs BLE między urządzeniami noszonymi a węzłami gwarantuje stabilność i energooszczędną pracę.

Takie podejście jest innowacyjne względem innych rozwiązań, np. systemu NITICS [2] czy niektórych komercyjnych systemów, przykładowo [8] (opaska noszona przez osobą starszą, parowana ze smartfonem).

Liczba węzłów stosowanych w systemie zależy przede wszystkim od rozmiaru obsługiwanego obszaru i jego specyfiki i nie można jej określić w sposób ogólny. Jednakże, można się pokusić o wstępne szacowanie, że dla przeciętnego mieszkania nie przekroczy ona liczby kilku sztuk.

3. STEROWNIK WĘZŁA SYSTEMU

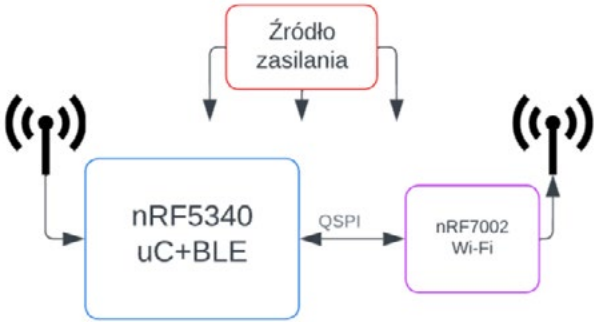
3.1. Ogólne założenia sterownika węzła

W systemie do monitorowania aktywności osób starszych bardzo ważną rolę odgrywa węzeł, który pośredniczy między jednostkowymi urządzeniami końcowymi, a zbiorczym urządzeniem bramki. Jego zadaniem jest zbieranie danych w formie pakietowej, rozgłaszanych protokołem BLE przez urządzenia pomiarowe (etykiety) oraz wysyłanie ich w odpowiedniej formie do węzła bramki za pomocą protokołu Wi-Fi.

Urządzenie węzła musi sprostać ściśle określonym wymaganiom czasowym podczas odbierania i analizy pakietów BLE. Potrzebny jest wówczas system czasu rzeczywistego, który obsługując mikrokontroler pozwoli na określenie ściśle ograniczonego czasu trwania procedury przetwarzania.

3.2. Schemat blokowy sterownika węzła

Na rys. 2 przedstawiono schemat węzła w postaci diagramu blokowego. Infrastruktura składa się z modułów radiowych (wraz z antenami), modułów przetwarzających dane (nRF5340 obsługującego BLE oraz kontrolera Wi-Fi nRF7002) oraz sekcji zasilania.



Rys. 2. Schemat blokowy modułu węzła

3.3. Opis modułów sterownika węzła

Moduł nRF5340 jest wielordzeniowym SoC (ang. System on Chip) zapewniającym nadzwyczajnie niski pobór prądu na poziomie 21,5mA w trybie pełnego obciążenia [10] oraz implementującym interfejsy bezprzewodowe. Jest to pierwszy tego typu układ zawierający jednocześnie dwa rdzenie Arm Cortex – M33 zdefiniowane jako: rdzeń sieciowy (Network Core) i rdzeń aplikacyjny (Application Core) [8]. Rdzeń sieciowy pozwala na integrację zasobów sieciowych – protokołu BLE wraz ze stosem i obsługą części radiowej, natomiast rdzeń aplikacyjny umożliwia równoczesne wykonywanie zadań zdefiniowanej aplikacji.

Moduł nRF7002 jest układem scalonym, który wdraża Wi-Fi 6 generacji w rozwiązaniach wymagających pracy w ograniczonym zakresie mocy. Implementuje on warstwę fizyczną (PHY) oraz warstwę kontroli dostępu do medium (MAC). Dzięki ograniczeniu do dwóch najniższych warstw protokołów modelu OSI, moduł ten jest wyjątkowo szybki i energooszczędny. Maksymalna szybkość transmisji jaką pozwala osiągnąć nRF7002 to 86 Mbps w warstwie fizycznej, która jednak jest ograniczona przepustowością interfejsu QSPI do 56 Mbps. Urządzenie pobiera maksymalnie 260 mA prądu podczas

nadawania sygnałów w paśmie 5 GHz z maksymalną mocą [11][12].

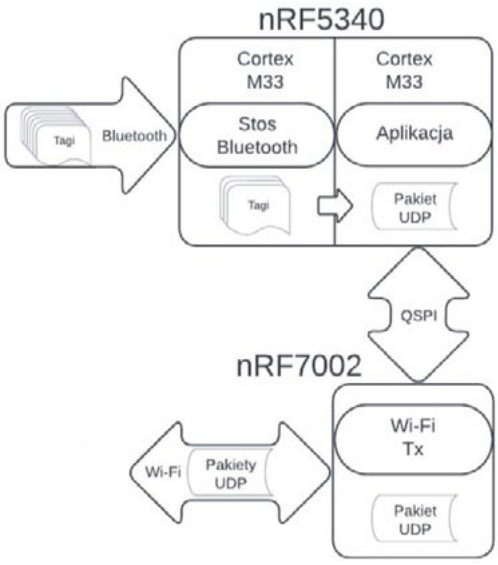
3.4. Wyzwania stawiane sterownikowi bramki

Dodatkową wartością wydaje się być potencjał wielordzeniowości, jaki zapewnia układ nRF5340. Jednoczesne przetwarzanie odebranych pakietów oraz nadawanie i odbieranie w dwóch różnych protokołach wymaga dużej złożoności obliczeniowej, czemu sprostać jest w stanie opisany SoC.

Na tym etapie nie ustalono jeszcze struktur danych przesyłanych w systemie, natomiast przyjęto założenie, że może być wiele urządzeń noszonych i wiele węzłów, co może generować znaczne ilości danych.

3.5. Schemat działania sterownika węzła

Podstawowym elementem węzła jest SoC nRF5340, obsługujący protokół BLE w ramach jednego z dwóch rdzeni Cortex M33. Drugi rdzeń nRF5340 umożliwia implementacje dowolnej aplikacji działającej równolegle ze stosem BLE [9], która w tym przypadku ma za zadanie przetworzyć pakiety odebrane w interfejsie BLE oraz stworzyć z nich zbiorczy pakiet UDP, który zostanie wysłany do bramki systemu, za pomocą protokołu Wi-Fi. Na rys. 3 jest przedstawiony schemat działania węzła.



Rys. 3 Schemat blokowy działania urządzenia

4. STEROWNIK BRAMKI SYSTEMU

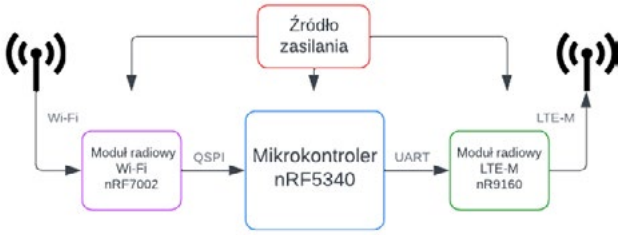
4.1. Ogólne założenia sterownika bramki

Równie ważna dla opisywanego systemu monitorowania aktywności osób starszych jest bramka (sterownik) systemu pośrednicząca w komunikacji pomiędzy sterownikiem węzła systemu a serwerem (bądź chmurą). Opisywana bramka ma za zadanie w pierwszej kolejności odbierać dane od sterownika węzła z wykorzystaniem interfejsu radiowego Wi-Fi. Następnie ma ona realizować wstępne przetwarzanie a także analizę danych z węzła systemu, przechowywanie ich na wypadek problemów z dalszym przesyłaniem oraz agregację pakietów. Ostatnim zadaniem bramki jest przesyłanie przetworzonych pakietów do serwera bądź chmury z wykorzystaniem sieci komórkowej.

W związku z postawionymi wymaganiami bramka musi w tym samym czasie komunikować się wykorzystując dwa typy interfejsów radiowych oraz przechowywać i przetwarzać przekazywane dane. Determinuje to wykorzystanie wydajnego mikrokontrolera i oprogramowanie go z wykorzystaniem systemu czasu rzeczywistego.

4.2. Schemat blokowy sterownika bramki

Na rys. 4 przedstawiono schemat blokowy sterownika bramki. Jego podstawowymi elementami są: moduły radiowe Wi-Fi – nRF7002 [11][12] i LTE-M – nRF9160 [13] oraz układ sterujący ich pracą oraz przetwarzający dane – mikrokontroler nRF5340 [9][10].



Rys. 4. Schemat blokowy modułu bramki

4.3. Opis modułów sterownika bramki

Moduł SIP (ang. System in Package) nRF9160 to moduł komunikacyjny firmy Nordic Semiconductor, zaprojektowany z myślą o zastosowaniach w dziedzinie Internetu Rzeczy (IoT) [4]. Moduł ten oparty jest na układzie, który łączy w sobie procesor ARM Cortex-M33 z rdzeniem ARMv8-M i moduł komunikacyjny LTE-M/NB-IoT zgodny z normami 3GPP LTE wersja 13 Cat-M1 i Cat-NB1 oraz 3GPP LTE wersja 14 Cat-NB1 i Cat-NB2 [7].

Zestaw urządzeń peryferyjnych oferuje różnorodne funkcje analogowe i cyfrowe, umożliwiające jednoukładową implementację szerokiej gamy aplikacji komórkowych IoT.

Modem LTE integruje bardzo elastyczny układ radiowy, który sprzętowo obsługuje zakres częstotliwości od 700 do 2200 MHz (poprzez pojedyncze złącze antenowe) oraz procesor pasma podstawowego, obsługujący warstwy L1-L3 protokołów LTE Cat-M1/NB1/NB2 oraz górne warstwy protokołu IP oferujące interfejs API dla aplikacji. Modem może być obsługiwany przez oprogramowanie Nordic Semiconductor, umożliwiające min. sterownie przy pomocy komend AT.

Moduły nRF7002 i nRF5340 zostały opisane bardziej szczegółowo sekcji 3.3.

4.4. Wyzwania stawiane sterownikowi bramki

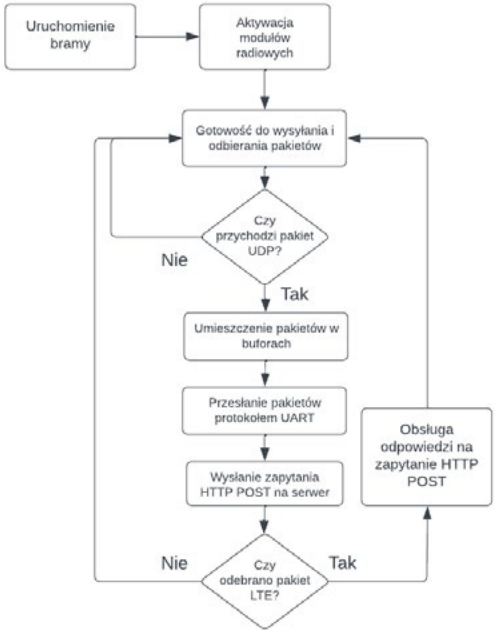
W kontekście zastosowania wymienionych modułów w sterowniku bramki systemu należy podkreślić ich cechy spójne ze stawianymi założeniami.

Moduły nRF7002 i nRF9160 implementują najnowsze standardy Wi-Fi 6 i LTE-M dzięki czemu zapewniają wysoką przepustowość łącza i perspektywę długoterminowej kompatybilności. Chociaż istnieje potencjalna możliwość wdrożenia technologii 5G, obecna powszechność i dostępność sieci LTE-M sprawia, że moduły te są preferowanym wyborem.

4.5. Schemat działania sterownika bramki

Na rys. 5 przedstawiono schemat działania braki od momentu jej uruchomienia.

W pierwszej kolejności po uruchomieniu sterownika bramki inicjalizowany jest moduł LTE w nRF9160. Po kilkunastu sekundach i nawiązaniu połączenia LTE urządzenie jest gotowe do działania. Węzły wysyłają pakiet UDP na adres IP modułu Wi-Fi nRF7002, będącego częścią bramki. W ramach modułu nRF7002 są one odbierane i selekcjonowane, a następnie przekazywane do mikrokontrolera nRF5340. Realizuje on wstępne przetworzenie i agregację pakietów. Części pakietów z danymi do wysłania są przekazywane do struktury komendy AT, mającej wysyłać zapytanie http POST na serwer. Utworzone w ten sposób wiadomości są umieszczane w buforach, które są wysyłane z wykorzystaniem protokołu UART do nRF9160, gdzie następuje realizacja komend AT i wysłanie zapytania.



Rys. 5 Schemat blokowy działania urządzenia

5. PODSUMOWANIE

W pracy przedstawiono ogólną architekturę systemu do monitorowania aktywności osób starszych. Centralną częścią takiego systemu jest jego infrastruktura złożona z węzłów odbierających dane od etykiet oraz bramki będącej stykiem między węzłami, a serwerem, bądź chmurą. Sterowniki węzłów oraz bramki powinny umożliwić wstępne przetwarzanie danych. Opracowana architektura i wykorzystanie najnowszych standardów powinny zapewnić elastyczność konfiguracji systemu i niezawodność jego działania.

Przedstawione sterowniki węzłów oraz bramki wyposażone są w najnowsze interfejsy radiowe Wi-Fi 6 i LTE-M oraz dwurdzeniowe mikrokontrolery mogące sprostać wszystkim stawianym im wymaganiom.

W referacie przedstawiono wyniki prac realizowanych w ramach projektu finansowanego przez Narodowe Centrum Badań i Rozwoju w ramach programu AAL Call 2021 nr umowy AAL/AALCall2021/52/CAREUP/2022 [1].

LITERATURA

[1] Active and Assisted Living Programme; CAREUP Project Description, dostęp 6.05.2023 http://www.aal-europe.eu/projects/careup/
[2] Active and Assisted Living Programme; NITICS Project Description, dostęp 6.05.2023, https://www.aal-europe.eu/projects/nitics/
[3] Akl, Ahmad, Babak Taati, i Alex Mihailidis. „Autonomous Unobtrusive Detection of Mild Cognitive Impairment in Older Adults". IEEE Transactions on Biomedical Engineering. Institute of Electrical and Electronics Engineers (IEEE), maj 2015. https://doi.org/10.1109/tbme.2015.2389149.
[4] Vishnubhatla, Arvind. 2020. "Cellular IOT Using nRF9160kit". International Journal of Online and Biomedical Engineering (iJOE) 16 (15):pp. 34-44. https://doi.org/10.3991/ijoe.v16i15.18987
[5] European Commission. The 2015 Ageing Report—Economic and Budgetary Projections for the 28 EU Member States (2013–2060); European Economy 3|2015; Directorate-General for Economic and Financial Affairs, Publications Office of the European Union: Brussels, Belgium, 2015.
[6] Chen, Datong, Ashok J. Bharucha, i Howard D. Wactlar. „Intelligent Video Monitoring to Improve Safety of Older Persons". 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, sierpień 2007. https://doi.org/10.1109/iembs.2007.4353163.
[7] Gabelle, Florian. "Narrowband-IOT power saving modes – A comprehensive study". Praca inżynierska. Oulu: Information Technology, Oulu University of Applied Sciences, 2021
[8] Opaska dla seniora, dostęp 20.06.2023, https://opaskadlaseniora.pl/
[9] nRF5340 – Nordic Semiconductor, dostęp 6.05.2023, https://www.nordicsemi.com/Products/nRF5340
[10] nRF5340 – Nordic Semicondutor Product Specification, dostęp 6.05.2023, https://infocenter.nordicsemi.com/pdf/nRF5340_PS_v1.3.pdf
[11] nRF7002 – Nordic Semiconductor, 6.05.2023, https://www.nordicsemi.com/Products/nRF7002
[12] nRF7002 – Nordic Semicondutor Product Specification, 6.05.2023, https://infocenter.nordicsemi.com/pdf/nRF7002_PS_v1.0.pdf
[13] nRF9160 – Nordic Semiconductor, dostęp 6.05.2023 https://www.nordicsemi.com/products/nrf9160
[14] Vuong, N. K., S. Chan, C. T. Lau, S. Y. W. Chan, P. L. K. Yap, i A. S. H. Chen. „Preliminary Results of Using Inertial Sensors to Detect Dementia- Related Wandering Patterns". 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, sierpień 2015. https://doi.org/10.1109/embc.2015.7319197.
[15] Strategy 2014–2020 for the Active and Assisted Living Programme, AAL, dostęp 6.05.2023, http://www.aal-europe.eu/wp-content/uploads/2015/11/20151001-AAL-Strategy_Final.pdf

SYSTEM DO BADANIA AKTYWNOŚCI OSÓB STARSZYCH Z WYKORZYSTANIEM PARAMETRÓW ŚRODOWISKOWYCH

SYSTEM FOR OLDER ADULTS' ACTIVITY INVESTIGATION BASED ON ENVIRONMENTAL PARAMETERS

Jakub Jędrys[1]; Jerzy Kołakowski[2]
[1] Instytut Radioelektroniki i Technik Multimedialnych, Politechnika Warszawska, Warszawa, jakub.jedrys.stud@pw.edu.pl
[2] Instytut Radioelektroniki i Technik Multimedialnych, Politechnika Warszawska, Warszawa, jerzy.kolakowski@pw.edu.pl

DOI: 10.15199/59.2023.4.81

Streszczenie: **W referacie przedstawiono koncepcję systemu umożliwiającego określenie bieżącej lokalizacji osoby starszej oraz identyfikację wykonywanych przez nią czynności poprzez obserwację zmian parametrów środowiskowych w pomieszczeniach, w których dana osoba przebywa. W tym celu węzły systemu lokalizacyjnego wyposażono w czujniki temperatury, wilgotności, ciśnienia atmosferycznego i stężenia gazów. W artykule zamieszczono opis koncepcji systemu oraz wyniki wstępnych badań zaproponowanego rozwiązania.**

Abstract: **The paper presents the concept of a system localizing older adults and detecting performed activities by observing changes in environmental parameters in rooms where the given person is staying. For this purpose, the localization system anchors were equipped with temperature, humidity, atmospheric pressure, and gas concentration sensors. The article contains a description of the system concept and the results of preliminary tests.**

Słowa kluczowe: **BLE, czujniki, IoT, monitorowanie osób starszych.**

Keywords: **BLE, sensors, IoT, older adults monitoring.**

1. WSTĘP

Według Światowej Organizacji Zdrowia (WHO) na utrzymywanie zdolności funkcjonalnych człowieka (functional abilities) wpływają przede wszystkim zdolności wewnętrzne (intrinsic capacity) oraz środowisko w jakim on żyje [1]. Zdolności te zmniejszają się wraz z wiekiem, przy czym zmiany te zachodzą szybciej w przypadku osób starszych. Określenie zdolności wewnętrznej osoby wymaga zbadania jej wydajność w każdej z pięciu określonych przez WHO dziedzin (mobilności, witalności, zdolności poznawczych, stanu psychicznego i stanu zmysłów wzroku i słuchu [2].

Badania zaprezentowane w referacie zostały zrealizowane w ramach projektu CAREUP, którego celem jest opracowanie platformy pozwalającej na kompensację zmian zdolności wewnętrznych wynikających z procesu starzenia. Koncepcja działania platformy polega na monitorowaniu osoby starszej, analizie wyników monitorowania oraz opracowaniu z pomocą grupy specjalistów (lekarzy różnych dyscyplin, specjalistów od opieki) i z udziałem osoby monitorowanej planu działań mających prowadzić do kompensacji zaobserwowanych zmian. Platforma gromadzi dane z obszarów: zdrowia fizycznego, funkcji poznawczych, doświadczeń egzystencjalnych, jakości życia, udziału w życiu społecznym i codziennego funkcjonowania. Wdrażanie i efekty planu działań są monitorowane, co pozwala na okresową aktualizację planu.

Monitorowanie osób starszych odbywa się z wykorzystaniem systemu IoT pozwalającego na rejestrację wyników z typowych urządzeń medycznych (m.in. z ciśnieniomierzy, wag, glukometrów), lokalizację osób w pomieszczeniach, rejestrację parametrów opisujących ruch osób. Dodatkowymi źródłami informacji są gry komputerowe, kwestionariusze i aplikacje działające na urządzeniach mobilnych.

Rozwiązanie zaproponowane w referacie rozszerza możliwości platformy poprzez dostarczenie informacji potwierdzających obecność osób w pomieszczeniach oraz wskazujących na wykonywanie przez te osoby dodatkowych czynności. Idea rozwiązania polega na umieszczeniu w węzłach systemu lokalizacyjnego dodatkowych czujników rejestrujących parametry środowiskowe.

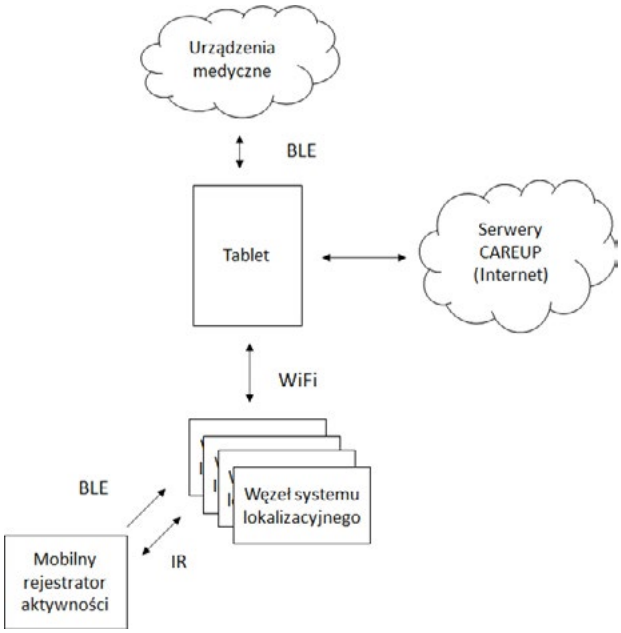
Tematyka wykorzystania parametrów środowiskowych do wyznaczania lokalizacji i detekcji aktywności była poruszana w nielicznych publikacjach. W artykule [3] autorzy opisują metodę polegającą na wykorzystaniu wartości wilgotności i temperatury rejestrowanych przez ruchomy węzeł do poprawy lokalizacji na podstawie rozkładu tych parametrów w budynku. Podobne rozwiązanie, lecz z użyciem czujników noszonych zaproponowano w [4]. Dodatkowo autorzy na podstawie wskazań czujników środowiskowych wykazali możliwość detekcji podstawowych czynności takich jak chodzenie i siedzenie. W [5] czujniki wilgotności i temperatury umieszczone w łazience zostały wykorzystane do detekcji mycia się. Z kolei w [6] opisano system złożony z kilkunastu czujników noszonych przez użytkownika (w tym czujników ciśnienia atmosferycznego), który umożliwia rozpoznanie kilkunastu codziennych czynności (ADL - activities of daily living). Zaproponowane w referacie rozwiązanie różni się od opisywanych w literaturze, ponieważ pomiary środowiskowe są wykonywane wyłącznie przez węzły systemu. Brak czujników środowiskowych w noszonym urządzeniu umożliwia zmniejszenie jego rozmiarów i energochłonności.

2. KONCEPCJA SYSTEMU

2.1. Ogólna architektura systemu

Na rys. 1 przedstawiono architekturę sieci IoT będącą elementem platformy CAREUP. Rolę bramki pełni tablet, którego zadaniem jest akwizycja wyników z pozostałych elementów systemu i transmisja do serwerów platformy. Tablet jednocześnie zapewnia interfejs użytkownika platformy, umożliwia prezentację wyników badań i planów działań oraz zapewnia dostęp do opracowanych na potrzeby platformy CAREUP gier dostarczających danych na temat kognitywnych możliwości osoby monitorowanej.

Użytkownik systemu jest wyposażony w mobilny rejestrator aktywności, przetwarzający wyniki pomiarów z czujników inercyjnych i wysyłający wyniki przetwarzania z użyciem interfejsu BLE (Bluetooth Low Energy) do tabletu za pośrednictwem węzłów systemu lokalizacyjnego. Rolą węzłów jest realizacja pomiarów, których wyniki służą do wyznaczania położenia osoby.

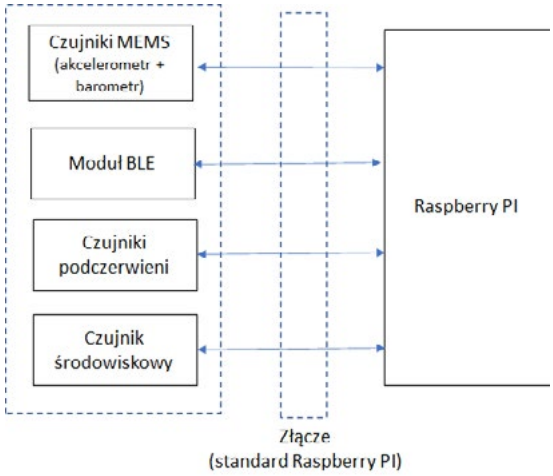


Rys. 1: Architektura sieci IoT w platformie CAREUP

2.2. Układ węzła systemu

Schemat blokowy węzła przedstawiono na rys. 2. Pracą węzła steruje minikomputer Raspberry Pi, wyposażony w nakładkę z czujnikami i modułami komunikacyjnymi. Moduł BLE, wyposażony w zewnętrzną antenę, jest wykorzystywany do komunikacji z etykietami i pomiaru poziomu sygnałów. W module dodatkowo zainstalowano czujniki podczerwieni. Czujnik obrazowy (kamerę termiczną) MLX90640 o rozdzielczości 32x24 piksele oraz laserowy czujnik zbliżeniowy o zasięgu 4m VL53L3CX. Fuzja danych z obu czujników z wynikami uzyskanymi z użyciem poziomów sygnałów BLE umożliwia zwiększenie dokładności lokalizacji.

Akcelerometr pozwala na wykrycie przemieszczenia węzła. Informacja ta jest istotna, ponieważ współrzędne węzła są uwzględniane w procesie wyznaczania lokalizacji. Barometr wykonuje pomiary referencyjne ciśnienia atmosferycznego, które stanowią odniesienie dla pomiarów wykonywanych w rejestratorach aktywności. Różnice ciśnień pozwalają na określenie wysokości, na której znajdują się etykiety względem wysokości zawieszenia węzła.



Rys. 2: Schemat węzła systemu

Pomiar warunków środowiskowych jest realizowany za pomocą układu BME680 firmy Bosch Sensortec. Rejestruje on parametry takie jak temperatura, wilgotność, ciśnienie i wykrywa stężenie gazów. Wartości odczytane z czujnika wymagają konwersji do użytecznych jednostek. Czujnik gazu jako surową wartość podaje opór odwrotnie proporcjonalny do stężenia lotnych związków organicznych w powietrzu. Przy zastosowaniu odpowiedniego algorytmu można na podstawie tej wartości obliczyć indeks jakości powietrza (IAQ) [8].

3. BADANIA SYSTEMU

3.1. Układ eksperymentalny

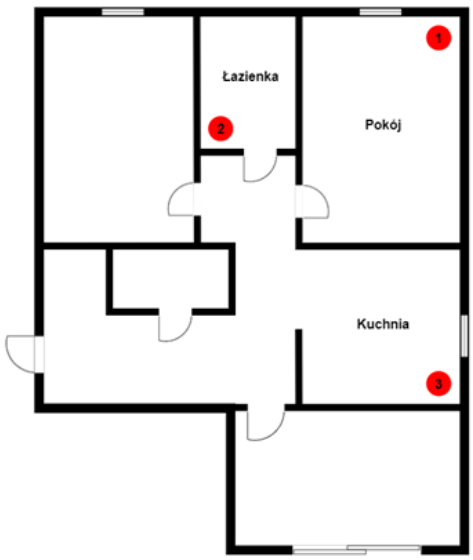
Celem badań była weryfikacja użyteczności wyników dostarczanych przez czujniki środowiskowe umieszczone w różnych pomieszczeniach mieszkania. Ponieważ cel nie zakładał konieczności wyznaczania lokalizacji, lecz jedynie długookresową rejestrację parametrów środowiskowych moduł z czujnikiem dołączono do energooszczędnego modułu mikrokontrolera nRF52833DK (Rys. 3). Wyniki pomiarów z czujnika BME680 były zapisywane do pliku tekstowego na karcie microSD.

Rys. 3: Schemat układu pomiarowego wykorzystanego podczas badań

Pracą układu steruje program działający w systemie operacyjnym czasu rzeczywistego Zephyr. Obsługa czujnika warunków środowiskowych jest realizowana w wydzielonym wątku. Wątek przechodzi w stan gotowości po

zadanym przedziale czasu (okres pomiarów) i pobiera surowe dane z czujnika. Następnie konwertuje surowe dane do użytecznych jednostek i zapisuje je w odpowiednich strukturach. Struktury te opisują zmiennoprzecinkowe wyniki pomiarów jako dwie wartości całkowite: część całkowitą i ułamkową danego pomiaru. Po skończonym pomiarze zmierzone wartości są kopiowane do message queue, a wątek przechodzi w stan nieaktywny. Po pojawieniu się danych w kolejce aktywuje się wątek odpowiedzialny za ich zapis do karty SD. Środowisko użyte do oprogramowania węzła to nRF Connect SDK for VS Code.

Do przeprowadzenia badań zostały użyte 3 układy rozmieszczone w charakterystycznych miejscach oznaczonych na planie mieszkania numerami od 1 do 3 (rys. 4). Pomiary były wykonywane z okresem 10 sekund.



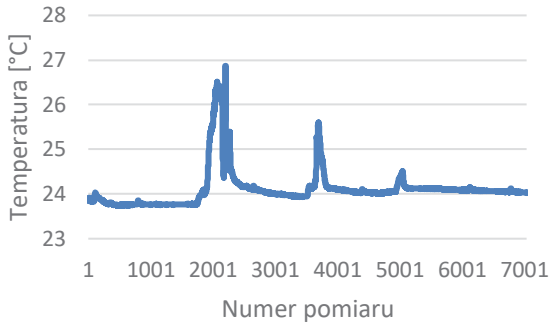
Rys. 4: Plan mieszkania z zaznaczoną lokalizacją układów testowych

3.2. Wstępne wyniki badań

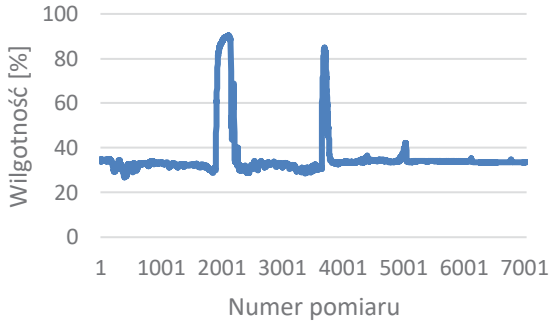
Na rysunkach od 5 do 10 przedstawiono przykładowe wykresy zmian zarejestrowanych parametrów w wybranych pomieszczeniach. Zrezygnowano z prezentacji wyników zgromadzonych w pokoju, ponieważ stwierdzono, że sama obecność osoby i realizowane przez nią typowe czynności nie powodują zauważalnych zmian parametrów. Wartości zewnętrznych czynników nie wpływają istotnie na wynik analizy danych ze względu na znaczną zmianę parametrów podczas badanych czynności.

Zmiany parametrów w łazience

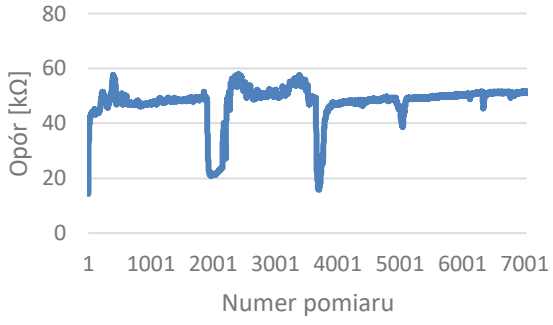
W trakcie trwania pomiarów w łazience zarejestrowano zmiany parametrów związane z kąpielą użytkowników pod prysznicem (w okolicach 2000 i 3700 pomiaru). Czynność ta spowodowała znaczny wzrost wartości temperatury (rys. 5) i wilgotności (rys. 6). Wykresy pozwalają na oszacowanie czasu trwania poszczególnych kąpieli. Zmiany parametrów zaobserwowano również na wykresie wskazań czujnika stężenia gazów. Gwałtowne zmiany oporu są jednak spowodowane zmianą wilgotności otoczenia i wynikają z zasady działania czujnika polegającej na pomiarze rezystancji podgrzewanej płytki pokrytej warstwą reagującą na obecność cząsteczek gazów.



Rys. 5: Zmiany temperatury w łazience



Rys. 6: Zmiany wilgotności w łazience



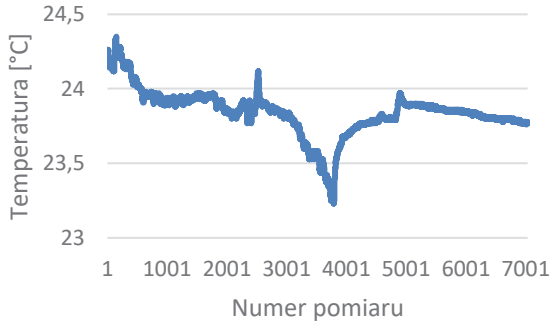
Rys. 7: Wskazania czujnika stężenia gazów w łazience

Zmiany parametrów w kuchni

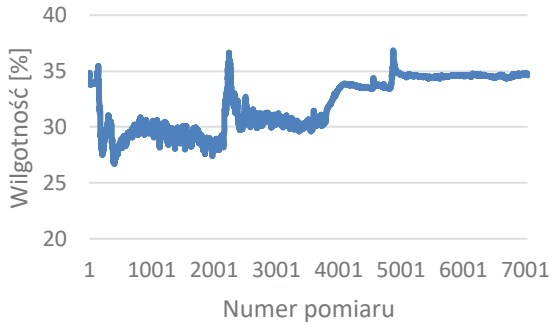
Czujniki umieszczone w kuchni umożliwiają wykrycie i oszacowanie czasu przygotowania posiłków. Gotowanie na kuchence w wykryto w okolicach 2500 pomiaru. W tym czasie zaobserwowano jednoczesny wzrost temperatury (rys. 8) i wilgotności (rys. 9) w pomieszczeniu. Udało się również zarejestrować czas, przez jaki było otwarte okno w kuchni (od początku pomiarów do okolic 3800 pomiaru). Zwłaszcza widoczny jest wyraźny wzrost

wartości oporu czujnika stężenia gazów (rys. 10) przy otwartym oknie.

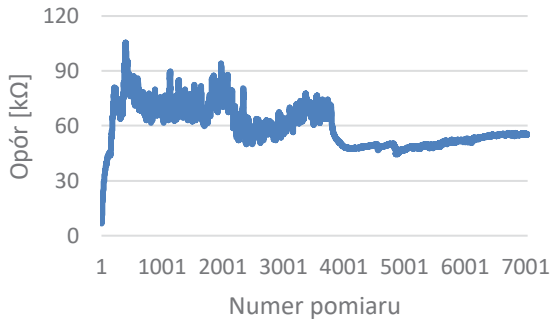
Wskazania czujnika w łazience lub kuchni pozwalają na stwierdzenie kontekstu obecności osoby w tych miejscach, jednakże oszacowanie czasu trwania czynności jest obciążone błędem wynikającym z miejsca instalacji czujnika oraz sprawności systemu wentylacji. W miejscu realizacji testów, w badanych pomieszczeniach była wykorzystana wymuszona wentylacja, co skutkowało dość szybkim powrotem parametrów środowiskowych do początkowych wartości. Badania potwierdziły ograniczoną przydatność czujnika stężenia gazów do pracy w środowisku o dużej wilgotności.



Rys. 8: Zmiany temperatury w kuchni



Rys. 9: Zmiany wilgotności w kuchni



Rys. 10: Wskazania czujnika stężenia gazów w kuchni

4. PODSUMOWANIE

W referacie przedstawiono koncepcję systemu pozwalającego na detekcję wybranych aktywności osób starszych poprzez obserwację parametrów środowiskowych (temperatury, wilgotności, ciśnienia, jakości powietrza) w pomieszczeniach, w których te czynności są przeprowadzane. System ten jest rozwinięciem systemu lokalizacyjnego wykorzystującego w procesie określania położenia sygnały systemu BLE.

Zaprezentowane wyniki wstępnych badań potwierdzają, że parametry środowiskowe mogą być źródłem informacji pozwalających na detekcję wybranych czynności np. przyrządzania posiłków, kąpieli, prania, jednak wykorzystanie tych danych wymaga wykonania wstępnych pomiarów, które pozwoliłyby określić te zmiany dla konkretnych miejsc instalacji czujników oraz w warunkach istniejącego systemu wentylacji pomieszczeń.

W referacie przedstawiono wyniki prac realizowanych w ramach projektu finansowanego przez Narodowe Centrum Badań i Rozwoju w ramach programu AAL Call 2021 nr umowy AAL/AALCall2021/52/CAREUP/2022.

LITERATURA

[1] Integrated care for older people (ICOPE): Guidance for person-centred assessment and pathways in primary care. Geneva: World Health Organization; 2019 (WHO/FWC/ALC/19.1).

[2] George, P.P., Lun, P., Ong, S.P. et al. A Rapid Review of the Measurement of Intrinsic Capacity in Older Adults. J Nutr Health Aging 25, 774–782 (2021).

[3] A. Rowe, Z. Starr and R. Rajkumar, "Using microclimate sensing to enhance RF localization in assisted living environments," 2007 IEEE International Conference on Systems, Man and Cybernetics, Montreal, QC, Canada, 2007, pp. 3668-3675, doi: 10.1109/ICSMC.2007.4414137.

[4] M. Jin, H. Zou, K. Weekly, R. Jia, A. M. Bayen and C. J. Spanos, "Environmental sensing by wearable device for indoor activity and location estimation," IECON 2014 - 40th Annual Conference of the IEEE Industrial Electronics Society, Dallas, TX, USA, 2014, pp. 5369-5375

[5] S. Chiriac, B. R. Saurer, G. Stummer and C. Kunze, "Introducing a low-cost ambient monitoring system for activity recognition," 2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops, Dublin, Ireland, 2011, pp. 340-345

[6] D. Roggen et al., "Collecting complex activity datasets in highly rich networked sensor environments," 2010 Seventh International Conference on Networked Sensing Systems (INSS), Kassel, Germany, 2010, pp. 233-240

[7] D. Oletic and V. Bilas, "Design of sensor node for air quality crowdsensing," 2015 IEEE Sensors Applications Symposium (SAS), Zadar, Croatia, 2015, pp. 1-5

[8] Bosch Sensortec„BME680 – Datasheet" BST-BME680-DS001-03

SYSTEM DO BADANIA CHODU OSÓB STARSZYCH

SYSTEM FOR OLDER ADULTS GAIT INVESTIGATION

Karol Duszczyk[1]; Jerzy Kołakowski[2]
[1] Instytut Radioelektroniki i Technik Multimedialnych, Politechnika Warszawska, Warszawa, karol.duszczyk.stud@pw.edu.pl
[2] Instytut Radioelektroniki i Technik Multimedialnych, Politechnika Warszawska, Warszawa, jerzy.kolakowski@pw.edu.pl

DOI: 10.15199/59.2023.4.82

Streszczenie: **W wyniku analizy parametrów chodu jest możliwa ocena aktywności oraz stanu zdrowia osoby starszej. W referacie opisano uniwersalny system rejestracji parametrów ruchu oparty na czujnikach przyspieszenia, prędkości kątowej i ciśnienia atmosferycznego. Zarejestrowane wyniki są transmitowane do bramki zlokalizowanej w budynku, w którym przebywa osoba, a po przetworzeniu przesyłane do serwera platformy PERHEART. Artykuł zawiera opis zrealizowanego układu oraz wstępne wyniki badań systemu rejestracji.**

Abstract: **Analysis of gait parameters proved to be useful in assessment of the activity and health of an older adult. The paper describes a gait parameter recording system utilizing acceleration, angular velocity, and atmospheric pressure sensors. The recorded results are transmitted to the gateway located in the building where the person is staying and sent to the PERHEART platform server after processing. The article describes the implemented system and preliminary results of the system tests.**

Słowa kluczowe: **analiza chodu, BLE, IoT**
Keywords: **BLE, gait analysis, IoT**

1. WSTĘP

Chód jest złożonym procesem motorycznym, który można opisać za pomocą wielu parametrów czasowo-przestrzennych (m.in. długości kroku, długości cyklu chodu, częstości kroków, czasów trwania pełnego cyklu chodu, prędkości) i kinematycznych (kątów w stawach, prędkości kątowych i liniowych) [1][2].

Chód jest wykonywany automatycznie przez osoby dorosłe. Wraz z wiekiem wykonywanie tej czynności staje się coraz trudniejsze, wymagające coraz większej kontroli motorycznej przez osobę starszą. Konsekwencją jej braku są upadki skutkujące poważnymi konsekwencjami. Analiza chodu staje się więc podstawowym narzędziem do oceny mobilności osób i ewentualnych zagrożeń upadkami [1].

Wyniki badań chodu osób starszych potwierdzają znaczne zmniejszenie się prędkości chodu, długości kroku oraz zwiększenie czasu wsparcia (czasu, gdy obie stopy mają kontakt z podłożem)[1]. Oprócz samych czynników związanych z wiekiem na chód wpływa również stan zdrowia a zwłaszcza choroby neurologiczne, układu kostno-szkieletowego [3], a także choroby serca [4][5]. Badania chodu pozwalają na ocenę ryzyka pojawienia się problemów zdrowotnych, a w niektórych przypadkach na detekcję choroby.

Większość badań chodu jest przeprowadzana w środowisku laboratoryjnym. Przykładem są badania opisane w [6] polegające na określeniu zmian chodu na dziesięciometrowej trasie lub podobne testy przeprowadzane z użyciem kamery [7]. W referacie [8] opisano system oparty na licznych czujnikach PIR (passive infrared) montowanych na sufitach pomieszczeń. Przykład systemów, w którym wykorzystano do tego celu noszone przez osoby czujniki inercyjne zaprezentowano w [9].Wyniki pomiarów są składowane na karcie microSD. W referatach [10][11] opisano rozwiązania bazujące na czujnikach umieszczanych w butach. Do transmisji wyników wykorzystano systemy ZigBee [10] i Bluetooth [11]. Autorom referatu nie są znane prace opisujące długoterminowe badania chodu w środowisku domowym, podczas codziennych czynności wykonywanych przez osoby monitorowane. System wykorzystywany do takich badań powinien charakteryzować się małą energochłonnością i małymi rozmiarami urządzeń.

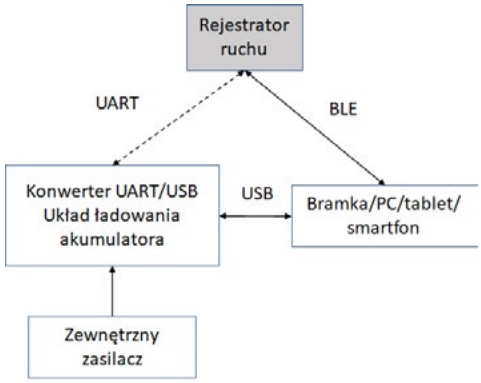
Przedstawione w referacie rozwiązanie jest elementem platformy sprzętowej opracowywanej w ramach projektu PERHEART, którego celem jest realizacja systemu wsparcia technologicznego w warunkach domowych dla pacjentów wypisywanych ze szpitala po hospitalizacji z rozpoznaniem niewydolności serca oraz dla pacjentów z niewydolnością serca korzystających z opieki ambulatoryjnej. Celem przeprowadzanych badań jest określenie zmian podstawowych parametrów chodu w cyklu dziennym i długoterminowym oraz korelacja tych zmian z wynikami uzyskanymi za pomocą standardowych urządzeń medycznych (m.in. wag, ciśnieniomierzy). W referacie opisano system służący do akwizycji wyników z czujników inercyjnych i miernika ciśnienia atmosferycznego. Dane te pozwolą na wyznaczenie podstawowych parametrów chodu w środowisku domowym, a także poza domem.

2. KONCEPCJA SYSTEMU

2.1. Ogólna architektura systemu

Na rys. 1 przedstawiono architekturę systemu. Dane pomiarowe są gromadzone przez rejestrator ruchu wyposażony w zestaw czujników.

Dane z rejestratora są przesyłane do urządzenia pełniącego rolę bramki IoT. W platformie PERHEART założono wykorzystanie do tego celu tabletu (umożliwia on również obsługę urządzeń medycznych wykorzystywanych w platformie), ale rolę tę może pełnić również komputer PC lub smartfon. Wykorzystanie przenośnej bramki w postaci smartfona umożliwia użycie systemu również podczas pobytu poza domem.

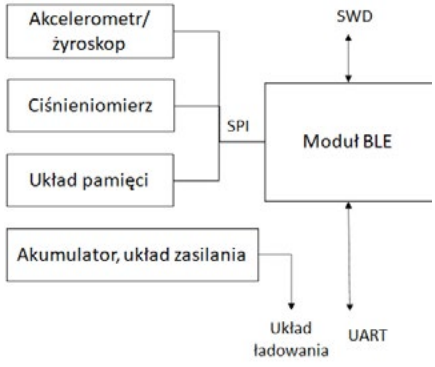


Rys. 1. Architektura systemu

Urządzenie może pracować w dwóch trybach. W pierwszym z nich wszystkie wyniki pomiarów są transmitowane do bramki, w drugim zaś wyniki pomiarów są składowane w lokalnej pamięci urządzenia, a część z nich (liczba kroków) jest transmitowana z użyciem łącza BLE. Odczyt wyników z wewnętrznej pamięci następuje po dołączeniu rejestratora do modułu wyposażonego w układ ładowania i konwerter łącza szeregowego do interfejsu USB. Zewnętrzny zasilacz jest jednocześnie układem zasilającym bramkę

2.2. Rejestrator parametrów ruchu

Schemat blokowy rejestratora zamieszczono na rys. 2.

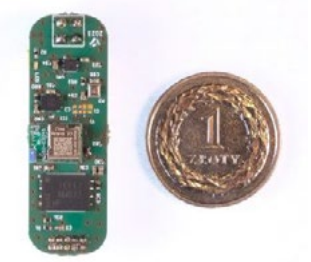


Rys. 2. Schemat blokowy rejestratora

W urządzeniu zostały zamontowane dwa układy MEMS firmy Bosch Sensortec. Akcelerometr BMI270 umożliwia rejestrację zmian przyspieszenia i prędkości kątowej, natomiast barometr BMP390 pozwala na wykrycie zmian wysokości na jakiej znajduje się układ. Oba układy dają możliwość kalibracji oraz przechowywania danych kalibracyjnych w swojej pamięci nieulotnej. Za sterowanie pracą rejestratora odpowiada moduł BL653u, w którym zastosowano mikrokontroler nRF52833. Mikrokontroler jest wyposażony w układ radiowy zapewniający komunikację w paśmie ISM 2.4 GHz. Rozwiązanie to umożliwia wykorzystanie różnych standardów sieci bezprzewodowych (m.in. BLE, ZigBee, Thread). Stos protokołów danej sieci jest wgrywany do pamięci układu wraz z oprogramowaniem sterującym pracą rejestratora.

Ponieważ nie można zagwarantować ciągłości komunikacji z bramką umieszczoną w pomieszczeniu, układ wyposażono w pamięć flash MT29F1G01ABAFDWB firmy Micron o pojemność 1 Gb. Przy próbkowaniu przyspieszenia i prędkości kątowej z częstotliwością 25 Hz, w ciągu sekundy otrzymywanych jest 300 bajtów danych (każda z próbek trójosiowych czujników inercyjnych zajmuje dwa bajty). Próbki ciśnienia są pobierane z okresem jednej sekundy, przez co nie powodują istotnego wzrostu objętości danych. Rozmiar pamięci umożliwia ciągły zapis wyników, przez ok. 120 godzin. W praktyce czas ten jest dłuższy, ponieważ układ wyposażono w mechanizm powodujący zatrzymanie akwizycji po 120 sekundowym okresie bezruchu.

Wszystkie układy peryferyjne są sterowane z użyciem interfejsu SPI, co zapewnia dużą szybkość transmisji. Jako źródło zasilania układu wykorzystano akumulator litowo-polimerowy. Ładowanie akumulatora jest możliwe po podłączeniu zewnętrznej ładowarki. Na rys. 3 przedstawiono fotografię zrealizowanego urządzenia.



Rys. 3. Fotografia płytki rejestratora

2.3. Oprogramowanie rejestratora

Oprogramowanie urządzenia zostało napisane w języku C z wykorzystaniem systemu operacyjnego czasu rzeczywistego Zephyr. Podczas tworzenia oprogramowania korzystano z Visual Studio Code z rozszerzeniem nRF Connect.

Do obsługi sensora BMI270 wykorzystano sterownik dostarczany z systemem Zephyr bez modyfikacji. Do obsługi sensora BMP390 wykorzystano sterownik do sensora BMP388, ponieważ sensory te mają identyczny układ rejestrów. W oryginalnym sterowniku należało tylko zmienić wartość identyfikatora układu z *0x50* na *0x60*.Do komunikacji Bluetooth wykorzystano Nordic UART Service (NUS). Jest to serwis Bluetooth, który pozwala na transparentne przekazywanie wiadomości w dwie strony.

W głównym programie zostały wydzielone 3 wątki:

- *bmp390_thread* - odpowiedzialny za inicjalizację i konfigurację sensora BMP390, odczyt pomiarów z sensora oraz przygotowanie danych do wysyłki
- *bmi270_thread* - odpowiedzialny za inicjalizację i konfigurację sensora BMI270, odczyt pomiarów z sensora oraz przygotowanie danych do wysyłki

- *ble_write_thread* - odpowiedzialny za odbieranie danych pomiarowych z pozostałych wątków oraz wysyłanie ich za pomocą Bluetooth

Wątki *bmp390_thread* i *bmi270_thread* przetwarzają surowe dane do postaci wiadomości, która zostanie przesłana do urządzenia odbiorczego:

- format wiadomości z sensora BMP390 to *1XXXXYY*, gdzie *XXXX* to 32-bitowa wartość ciśnienia w mPa (4 wartości 8-bitowe), *YY* to 16-bitowa wartość temperatury w centy Kelvinach (2 wartości 8-bitowe)
- format wiadomości z sensora BMP390 to *2KKLLMMNNOOPP*, gdzie *KK,LL,MM* to wartości pomiarów z kolejnych osi akcelerometru (x, y, z), *NN,OO,PP* to wartości pomiarów z kolejnych osi żyroskopu (x, y, z). Wartości zostały zamienione na liczby całkowite $\left(\frac{1}{1000}\frac{\mathrm{m}}{\mathrm{s}^2}, \frac{1}{1000}\frac{\mathrm{rad}}{\mathrm{s}}\right)$ oraz powiększone o 32768 (2^{15}) w celu wyeliminowania liczb ujemnych

Następnie wątki umieszczają sformatowaną wiadomość w kolejce FIFO. Wątek *ble_write_thread* odbiera kolejno wiadomości z kolejki FIFO i wysyła je za pomocą Bluetooth.

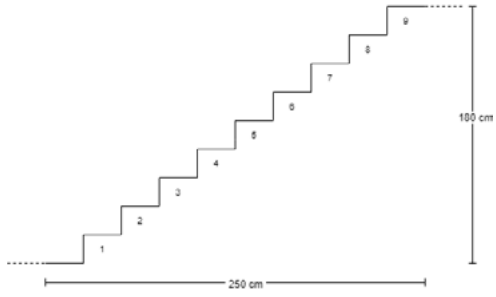
3. BADANIA SYSTEMU

3.1. Układ eksperymentalny

Badania przeprowadzono w warunkach laboratoryjnych. Układ eksperymentalny składał się z rejestratora, który wykonywał pomiary z częstotliwością 10Hz dla czujnika BMP390 oraz 25Hz dla czujnika BMI270 oraz telefonu z systemem Android z zainstalowaną aplikacją *Serial Bluetooth Terminal*. Następnie zebrane dane były przeniesione na komputer i przetwarzane w języku Python z wykorzystaniem bibliotek Pandas i Matplotlib.

W celu zbadania systemu przetestowano 2 typowe scenariusze:

1. wejście po schodach [rys. 4], obrót o 180°, zejście ze schodów
2. chód po linii prostej o długości 650 cm, obrót o 180°, bieg po linii prostej o długości 650 cm



Rys. 4. Schemat schodów

Dla każdego scenariusza przetestowano 2 ułożenia urządzenia: urządzenie przymocowane do paska osoby badanej, urządzenie przymocowane do prawego buta osoby badanej. Docelowo urządzenie będzie umieszczane w obu butach w celu wykrywania asymetrii chodu.

Podczas analizowania zebranych danych obliczono wysokość (względem wysokości początkowej), na której znajdowało się urządzenie w danym czasie. Wykorzystano do tego wzór barometryczny:

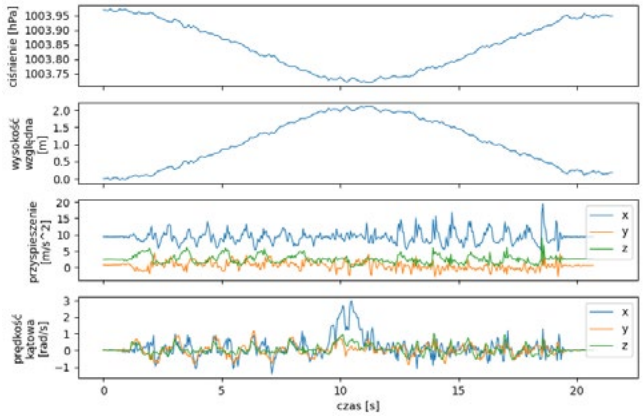
$$p = p_0 \exp\left(-\frac{\mu g h}{RT}\right), \tag{1}$$

gdzie: p_0 - ciśnienie atmosferyczne na poziomie odniesienia (z pierwszej próbki), μ - masa molowa powietrza ($0{,}0289644\ \frac{kg}{mol}$), g - przyspieszenie ziemskie, R - stała gazowa ($8{,}314463\ \frac{J}{mol \cdot K}$), T - temperatura powietrza w K (dla uproszczenia - średnia dla wszystkich pomiarów)

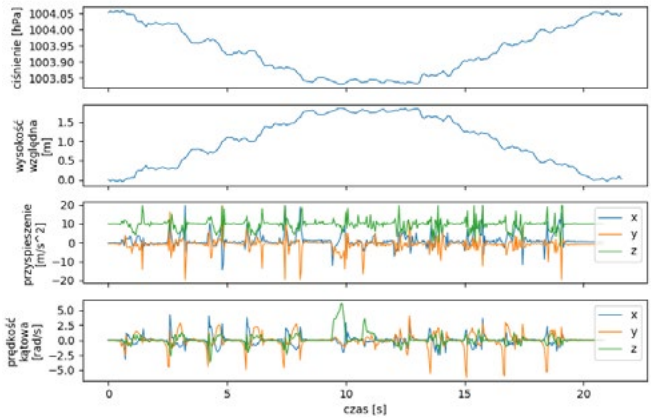
Po przekształceniu równania 1 otrzymano wzór na wysokość względną:

$$h = -\frac{RT}{\mu g} \ln\left(\frac{p}{p_0}\right) \tag{2}$$

3.2. Wyniki badań



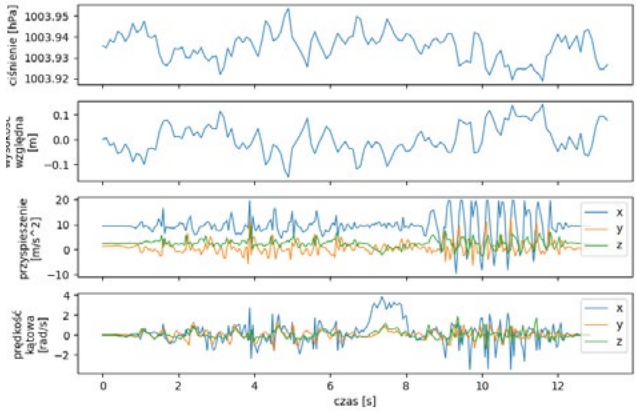
Rys. 5. Schody - urządzenie na pasku



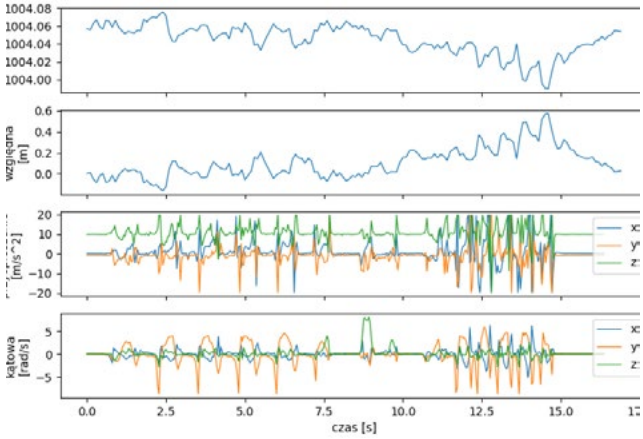
Rys. 6. Schody - urządzenie na bucie

W obu badaniach scenariusza ze schodami udało się zaobserwować zmianę wysokości względnej (związanej ze zmianą ciśnienia atmosferycznego). Podczas wchodzenia po schodach wysokość rosła, a podczas schodzenia z nich malała. Gdy urządzenie było zamocowane do paska osoby badanej urządzenie wskazywało dla każdego stopnia schodów niewielkie zmiany wartości przyspieszenia i prędkości kątowej. Ze wskazań żyroskopu można wywnioskować, że koło 10-tej sekundy nastąpił obrót o 180°. Gdy urządzenie było zamocowane do prawego buta osoby badanej urządzenie wskazywało zmiany przyspieszenia i prędkości kątowej około 2 razy rzadziej. Wiąże się to z tym, że prawa stopa była stawiana jedynie na co drugim stopniu (stopnie 1, 3, 5, 7, 9), a na pozostałych stopniach

stawiana była lewa stopa. Zmiana wysokości w tym przypadku jest mniej płynna (skokowe zmiany wartości).



Rys. 7. Chód - urządzenie na pasku



Rys. 8. Chód - urządzenie na bucie

Podczas chodu z urządzeniem zamocowanym na pasku osoby badanej zmiany wysokości były niewielkie i wynikały prawdopodobnie z niedokładności pomiaru i lekko niestabilnego chodu. Zmiany prędkości kątowej i przyspieszenia wynikają z dynamiki chodu i są największe podczas biegu. Ze wskazań żyroskopu można wywnioskować, że około 7-8 sekundy nastąpił obrót o 180°. Gdy urządzenie było zamocowane do buta można zaobserwować większe wahania wysokości względnej wynikające z podnoszenia i opuszczania stopy. Na wykresie widać gwałtowne zmiany przyspieszenia i prędkości kątowej dla każdego stawianego kroku.

4. PODSUMOWANIE

W referacie przedstawiono system do rejestracji parametrów chodu opracowany na potrzeby projektu PERHEART. W systemie wykorzystano czujniki inercyjne (akcelerometr i żyroskop) do pomiaru parametrów ruchu oraz czujnik barometrycznych do rejestracji zmian wysokości. Wyniki pomiarów są wysyłane do bramki systemu z wykorzystaniem interfejsu BLE. Mały rozmiar wykonanego czujnika pozwala na jego zamocowanie zarówno w obuwiu jak i wszycie w ubranie osoby monitorowanej. Przeprowadzone wstępne testy systemu potwierdziły skuteczność zaproponowanego rozwiązania.

W referacie przedstawiono wyniki prac realizowanych w ramach projektu finansowanego przez Narodowe Centrum Badań i Rozwoju w ramach programu EraPerMed nr umowy PerMed/II/34/PerHeart/2022.

LITERATURA

[1] Bridenbaugh S.A., Kressig R. W., “Laboratory Review: The Role of Gait Analysis in Seniors’ Mobility and Fall Prevention”. Gerontology 2011;57:256–264, DOI: 10.1159/000322194

[2] Middleton A., Stacy F. L., (2013). Assessment of Gait, Balance, and Mobility in Older Adults: Considerations for Clinicians. Current Translational Geriatrics and Experimental Gerontology Reports. 2. 205-214. 10.1007/s13670-013-0057-2.

[3] Salzman B. Gait and balance disorders in older adults. Am Fam Physician. 2010 Jul 1;82(1):61-8. PMID: 20590073.

[4] Ozcan EB, Saglam M, Vardar-Yagli N, Calik-Kutukcu E, Inal-Ince D, Altinsoy M, Kaya EB. “Impaired Balance and Gait Characteristics in Patients With Chronic Heart Failure”. Heart Lung Circ. 2022 Jun;31(6):832-840. Epub 2022 Feb 15.

[5] Chen MA. Frailty and cardiovascular disease: potential role of gait speed in surgical risk stratification in older adults. J Geriatr Cardiol. 2015 Jan;12(1):44-56. PMID: 25678904; PMCID: PMC4308458.

[6] C. Soaz and K. Diepold, "Step Detection and Parameterization for Gait Assessment Using a Single Waist-Worn Accelerometer," in IEEE Transactions on Biomedical Engineering, vol. 63, no. 5, pp. 933-942, May 2016,

[7] X. Cai, G. Han, X. Song and J. Wang, "Single-Camera-Based Method for Step Length Symmetry Measurement in Unconstrained Elderly Home Monitoring," in IEEE Transactions on Biomedical Engineering, vol. 64, no. 11, pp. 2618-2627, Nov. 2017,

[8] S. Hagler, D. Austin, T. L. Hayes, J. Kaye and M. Pavel, "Unobtrusive and Ubiquitous In-Home Monitoring: A Methodology for Continuous Assessment of Gait Velocity in Elders," in IEEE Transactions on Biomedical Engineering, vol. 57, no. 4, pp. 813-820, April 2010, doi: 10.1109/TBME.2009.2036732.

[9] N. Shiozawa, S. Okada and M. Makikawa, "Gait assessment for elderly using a portable acceleration monitoring device", Proceedings of SICE Annual Conference 2010, Taipei, Taiwan, 2010, pp. 2538-2539.

[10] T. Nilpanapan and T. Kerdcharoen, "Social data shoes for gait monitoring of elderly people in smart home," 2016 9th Biomedical Engineering International Conference (BMEiCON), Laung Prabang, Laos, 2016, pp. 1-5,

[11] M. I. Mohamed Refai, B. -J. F. van Beijnum, J. H. Buurke and P. H. Veltink, "Gait and Dynamic Balance Sensing Using Wearable Foot Sensors," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 27, no. 2, pp. 218-227, Feb. 2019.

KOD FONTANNOWY Z PRZYRASTAJĄCĄ LICZBĄ SYMBOLI ŹRÓDŁOWYCH

FOUNTAIN CODE WITH INCREASING NUMBER OF SOURCE SYMBOLS

Kamila Haustein[1]; Marek Blok [2]

[1] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, hausteinkamila@gmail.com
[2] Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Gdańsk, marek.blok@pg.edu.pl

DOI: 10.15199/59.2023.4.83

Streszczenie: **Kody fontannowe zabezpieczające transmisję przed wymazaniami wyróżnia brak określonej z góry długości i losowy sposób generacji kolejnych pakietów kodowych. W pracy pokazano jak właściwości tych kodów zależą od rozkładu macierzy generującej dla dwóch wariantów dostarczania pakietów do kodera. Szczególnie interesujący jest przypadek, gdy pakiety źródłowe trafiają do kodera stopniowo podczas transmisji. Uzyskane wyniki wskazują na możliwość realizacji transmisji strumieniowej o ograniczonej złożoności i wymaganiach pamięciowych dekodera.**

Abstract: **Fountain codes that protect the transmission against erasures are distinguished by the lack of a predetermined length and the random method of generation of successive code packets. The paper shows how the properties of these codes depend on the distribution of the generating matrix for two variants of packet delivery to the encoder. Of particular interest is the case where the source packets reach the encoder gradually during transmission. The obtained results indicate the possibility of stream transmission with limited complexity and memory requirements of the decoder.**

Słowa kluczowe: **nadmiarowe kodowanie korekcyjne, kanał z wymazywaniem, kody fontannowe, transmisja strumieniowa.**

Keywords: **erasure channel, error correction coding, fountain codes, stream transmission.**

1. WSTĘP

W dzisiejszych czasach do przesyłania wiadomości korzystamy z technologii takich jak Internet czy telefon komórkowy. Podczas transmisji przez kanał mogą występować błędy, które wynikają ze zniekształceń i zakłóceń. Powoduje to, że część pakietów koniecznych do przesłania zostaje wymazana i utracona, co powoduje, że nie dotrą one do dekodera. Poprzez zastosowanie kodów fontannowych mamy możliwość odtworzenia wymazanych pakietów i poprawne przesłanie wiadomości bez względu na warunki panujące w kanale. Kody te nie mają ograniczonej długości i nadawane pakiety kodowe generowane są aż do momentu, w którym dekoder jest w stanie odzyskać przekazywane dane źródłowe.

Koncepcja kodów fontannowych zakłada, że informacja o pakietach źródłowych jest rozpraszana w wielu pakietach kodowych, tak żeby po stronie odbiorczej możliwe było prawidłowe odtworzenie wiadomości źródłowej na podstawie jak najmniejszego podzbioru nadawanych pakietów pomimo losowej utraty niektórych pakietów podczas transmisji. Istnieje kilka rodzajów kodów fontannowych takich jak kody Tornado [4], kody LT [4] oraz kody Raptor [4] i ich usprawniona wersja kody RaptorQ [7].

Podstawową cechą kodów fontannowych jest, to że bieżący symbol kodowy jest generowany jako suma modulo dwa losowo wybranych symboli źródłowych. Z kolei po stronie odbiorczej odtworzenie wybranego symbolu źródłowego wymaga odnalezienia takiego zestawu symboli kodowych, których suma modulo dwa jest równa odtwarzanemu symbolowi źródłowemu. Taka konstrukcja oznacza, że po stronie nadawczej zawsze możemy wygenerować kolejny symbol kodowy, a po stronie nadawczej ma znaczenie przede wszystkim, ile symboli kodowych udało się odebrać, a niektóre z nich zostały wymazane w kanale. Koncepcję tę łatwo przenieść z kodowania symboli dla kanału z wymazywaniem do kodowania pakietów dla kanału z wymazywaniem pakietów.

Chociaż kody fontannowe gwarantują, że przesyłana wiadomość zostanie poprawnie odebrana niezależnie od prawdopodobieństwa wymazywania, to dobór sposobu losowego generowania kolumny macierzy generującej kodu wskazującej pakiety źródłowe tworzące pakiet kodowy wpływa na liczbę pakietów kodowych potrzebnych do zdekodowania przekazywanej wiadomości oraz opóźnienia dekodowania poszczególnych jej pakietów. Przeprowadzone badania miały na celu zweryfikowanie wpływu macierzy generującej na pracę dekodera dla dwóch wariantów: macierzy o równomiernym rozkładzie jedynek oraz macierzy o wagach kolumn losowanych zgodnie z rozkładem solitonowym [8].

Kolejny analizowany w pracy aspekt, dotyczy sposobu dostarczania danych do kodera w nadajniku. Podstawowa koncepcja kodów fontannowych zakłada, że koder w nadajniku zna od razu wszystkie przekazywane symbole/pakiety źródłowe. W pracy rozpatrujemy dodatkowo wariant, w którym kolejne pakiety źródłowe trafiają do kodera stopniowo i koder generuje pakiety kodowe na bieżąco jeszcze za nim otrzyma całą wiadomość do przekazania. Eksperymenty te miały na celu określenie jak zmiana strategii wpływa na liczbę pakietów wymaganych w dekoderze do zdekodowania całej wiadomości oraz opóźnienia dekodowania jej poszczególnych pakietów. Badania te są wstępnym krokiem w kierunku opracowania wariantu kodu fontannowego pozwalającego na strumieniową transmisję danych.

2. KODY RAPTOR

Kody Raptor zostały opatentowane w 2001 roku i są one najnowocześniejszym rozwiązaniem kodów fontannowych [4]. Są one rozszerzeniem kodów LT i umożliwiają odebranie danych nadawanych przy mniejszych wymaganiach. Różnią się tym, że nie wymagają one odzyskania wszystkich pakietów źródłowych, aby zdekodować całą wiadomość.

Kod Raptor składa się z kodera wstępnego oraz właściwego kodera LT. Kod wstępny (*pre-coder*) jest standardowych kodem blokowym, który ma za zadanie dodatkowo zabezpieczyć kodowaną wiadomość przed wymazaniami i pozwala on na odtworzenie części brakujących pakietów otrzymywanych na wyjściu dekodera wewnętrznego kodu LT. Zauważmy, że w przypadku kodu LT im mniejsza jest liczba jeszcze niezdekodowanych pakietów tym bardziej spada prawdopodobieństwo, że kolejny odebrany pakiet kodowy pozwoli na odtworzenie jednego z brakujących pakietów źródłowych. Oznacza, to że możliwość odtworzenia kilku ostatnich pakietów źródłowych z wykorzystaniem kodu wstępnego przekłada się na wcześniejsze zdekodowanie wiadomości. Jednocześnie pozwala to na zmniejszenie złożoności algorytmu dekodowania kodu LT poprzez użycie macierzy generującej o niższych wagach kolumn, co jednocześnie zmniejsza prawdopodobieństwo zdekodowania końcowych pakietów źródłowych.

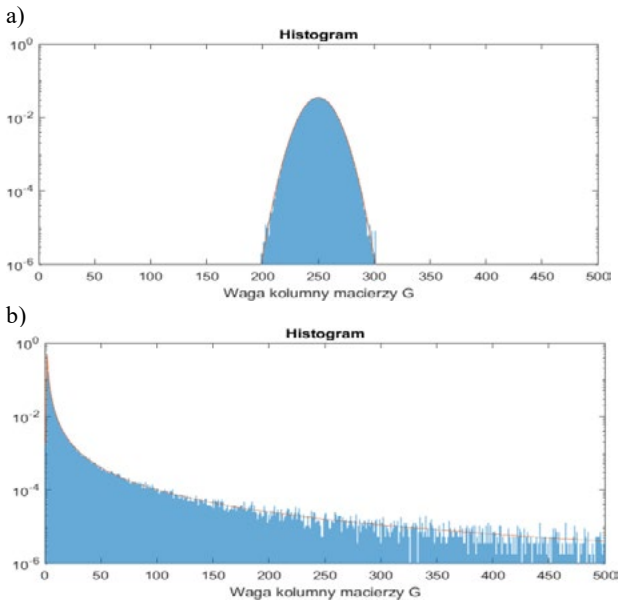
Najnowocześniejszym i najbardziej elastycznym oraz wydajnym rodzajem kodów Raptor są kody RaptorQ [7]. Kody te oferują rozwiązania spełniające wymagania zarówno użytkowników końcowych, jak również wysokiej jakości dostawców usług (QoS). Koder oraz dekoder pozwala usługom na przesyłanie strumieniowe, odzyskanie danych utraconych w trakcie transmisji oraz ich rekonstrukcję bez użycia kanału zwrotnego. Blok źródłowy jest konfigurowany przez aplikację i jest on podzielony na równe fragmenty danych. Koder RaptorQ generuje pakiety naprawy (*repair symbols*) za pomocą pakietów źródłowych bloku. Natomiast zaszyfrowane pakiety, które są kombinacją pakietów źródłowych i naprawczych, są wysyłane do odbiorcy.

3. BADANIE WŁAŚCIWOŚCI WYBRANYCH WARIANTÓW KODÓW FONTANNOWYCH

W ramach eksperymentów zbadano właściwości wybranych wariantów kodów fontannowych pozwalających na odzyskanie danych wymazanych w trakcie transmisji bez użycia kanału zwrotnego. Porównano tu efekty dla zastosowania dwóch rozkładów wag kolumn macierzy generującej (rys. 1). Wyniki wygenerowano na podstawie symulacji transmisji 1000 wiadomości złożonych z 500 pakietów źródłowych.

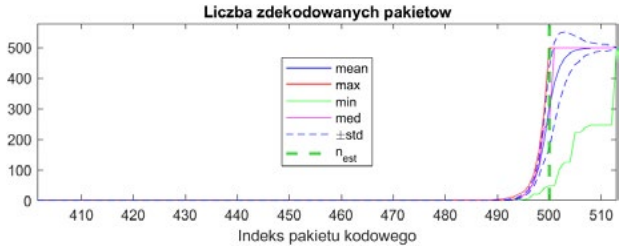
3.1. Scenariusz referencyjny

W niniejszym punkcie założono transmisję przez kanał idealny a wszystkie pakiety źródłowe trafiają od razu do kodera w nadajniku, co jest podstawowym założeniem dla kodów fontannowych.



Rys. 1. Histogram wag macierzy generującej dla (a) rozkładu dwumianowego oraz (b) rozkładu solitonowego

Na rys. 2 przedstawiono wyniki uzyskane dla równomiernego rozkładu jedynek w macierzy generującej G. Założono, że jedynka na danej pozycji występuje z prawdopodobieństwem 0.5, z wyłączeniem przypadku, w którym kolumna macierzy G składa się z samych zer. Przekłada się to na dwumianowy rozkład wag kolumn macierzy generującej o wartości średniej równej połowie liczby przesyłanych pakietów (rys. 1a). Dla uproszczenia dalej ten przypadek macierzy G będzie nazywany macierzą o rozkładzie dwumianowym. Zauważmy, że w tym przypadku do blisko 500 iteracji nie został zdekodowany żaden pakiet, a następnie widoczny jest gwałtowny wzrost liczby zdekodowanych pakietów i po otrzymaniu kilku dodatkowych pakietów cała wiadomość zostaje poprawnie zdekodowana.



Rys. 2. Liczba zdekodowanych pakietów w funkcji indeksu odebranego pakietu kodowego dla transmisji przez kanał idealny dla rozkładu dwumianowego

Z kolei dla solitonowego rozkładu wag kolumn macierzy generującej (rys. 3) możemy zauważyć, że pierwsze pakiety są odtwarzane po odebraniu średnio około 350 pakietów kodowych (rys. 2) chociaż, w niektórych przypadkach proces dekodowania początkowych pakietów zaczyna się już po około 100 iteracjach, Gwałtowny wzrost liczby zdekodowanych pakietów widoczny jest dla około 480 iteracji, a następnie proces ten spowolnia ponieważ maleje prawdopodobieństwo, że kolejny pakiet odebrany pozwoli na zdekodowane kolejnego pakietu źródłowego. Cała wiadomość zostaje odebrana w połowie przypadków po około 550 pakietach kodowych, ale w najgorszym przypadku wymagane było odebranie aż 1000 pakietów kodowych.

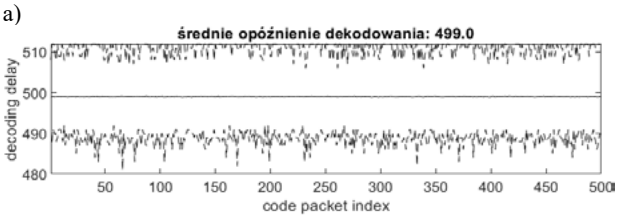
Na podstawie przedstawionych danych oraz wykresów widać, że rozkład solitonowy pomimo tego, że pozwala na wcześniejsze zdekodowanie pierwszych pakietów to równocześnie wymaga większej liczby przesłanych i odebranych pakietów zwiększając opóźnienie wynikające z kodowania.

Rys. 3. Liczba zdekodowanych pakietów w funkcji indeksu odebranego pakietu kodowego dla transmisji przez kanał idealny dla rozkładu solitonowego

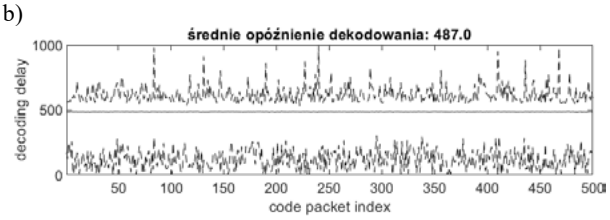
Potwierdza to opóźnienie dekodowania poszczególnych pakietów źródłowych (rys. 4). Dla rozkładu o dużej wartości średniej (dwumianowego) opóźnienie dekodowania wykazuje relatywnie nieduże wahania. Minimalne zaobserwowane opóźnienie wskazuje, kiedy najwcześniej pierwszy pakiet źródłowy może dostać zdekodowany, a maksymalne zaobserwowane opóźnienie wskazuje na to, kiedy ostatni pakiet źródłowy, kończący dekodowanie całej wiadomości może zostać zdekodowany. Dla rozkładu dwumianowego (rys. 4a) widzimy, że w ramach 1000 symulowanych transmisji najwcześniej po 481 odebranych pakietach kodowych pojawił się pierwszy zdekodowany pakiet, jednocześnie ostatni pakiet został zdekodowany nie później niż po 512 pakietach kodowych.

W przypadku rozkładu solitonowego (rys. 4b) chociaż obserwujemy mniejsze średnie opóźnienie dekodowania (487 zamiast 499), to jednocześnie zakres wahań opóźnienia dekodowania poszczególnych pakietów znacznie wzrasta. Wartość minimalna tym razem wynosi nawet 0 (pierwszy pakiet kodowy może być równoważny jednemu z pakietów źródłowych), ale maksymalna sięga aż 1000.

a)



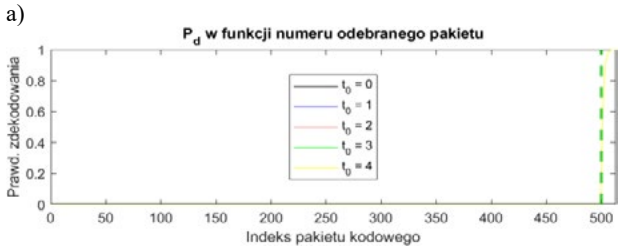
b)



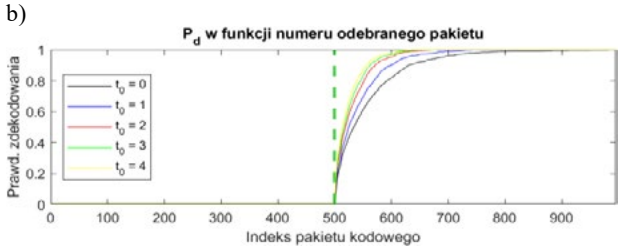
Rys.4. Opóźnienie dekodowania pakietów źródłowych dla kanału idealnego oraz rozkładu (a) dwumianowego i (b) solitonowego; liniami przerywanymi zaznaczono wartość minimalną i maksymalną wyznaczoną na podstawie 1000 transmisji

Porównując wartości prawdopodobieństwa zdekodowania wiadomości P_d w funkcji liczby pakietów odebranych dla rozkładu dwumianowego i solitonowego (rys. 5 oraz tab. 1) możemy zauważyć, że dla rozkładu dwumianowego bez względu na wartość t_0 (oznaczającą liczbę pakietów źródłowych odtwarzanych przed koder wstępny) liczba wymaganych pakietów kodowych jest taka sama. Oznacza to, że w tym przypadku zastosowanie kodera wstępnego mija się z celem. Natomiast, dla rozkładu solitonowego wraz ze wzrostem wartości t_0 potrzebna jest coraz mniejsza liczba pakietów, co uzasadnia stosowanie w tym przypadku kodera wstępnego.

a)



b)



Rys. 5. Prawdopodobieństwo zdekodowania dla kanału idealnego oraz rozkładu (a) dwumianowego i (b) solitonowego

Tab.1. Liczba przesłanych pakietów kodowych wymaganych dla osiągnięcia wybranych wartości P_d

P_d	t_0	rozkład dwumianowy	rozkład solitonowy
1,0	0	513	990
	1		975
	2		691
	3		682
	4		680
0,95	0	505	666
	1		621
	2		593
	3		578
	4		570

3.2. Właściwości transmisji dla stopniowego napływania pakietów źródłowych

Rozpatrzmy teraz przypadek transmisji przez idealny kanał jednokierunkowy bez strat pakietów, w którym pakiety źródłowe stopniowo wpływają do nadajnika w trakcie przesyłania pakietów kodowych. Zakładamy tu, że pakiety kodowe generowane są w tym samym tempie jak napływają do kodera pakiety źródłowe, a maksymalna liczba pakietów źródłowych wynosi 500. Zauważmy, że zmienia to również rozkłady wag kolumn macierzy generującej, bo długość kolumny macierzy generującej rośnie wraz z przyrostem liczby pakietów źródłowych a wraz z nią rośnie średnia liczba jedynej lokowanych w kolumnie.

Rys. 6 przedstawia statystyki liczby zdekodowanych pakietów w funkcji indeksu pakietu kodowego pomierzone dla rozkładu dwumianowego. Jak widać tym razem pierwsze pakiety są dekodowane są w przeciągu całej transmisji i średnio na dwa otrzymane pakiety zdekodowany zostaje jeden nowy pakiet źródłowy (rys. 6). Średnio po otrzymaniu pakietów kodowych w liczbie równej maksymalnej liczbie pakietów źródłowych zostaje zdekodowana połowa pakietów. Następnie po otrzy-

maniu kilku dodatkowych pakietów liczba ta gwałtownie wzrasta i cała wiadomość zostaje zdekodowana, chociaż nieco później niż dla wcześniejszego przypadku (tab. 1 i tab. 2).

Rys. 6. Liczba zdekodowanych pakietów dla rozkładu dwumianowego i pakietów napływających do kodera stopniowo

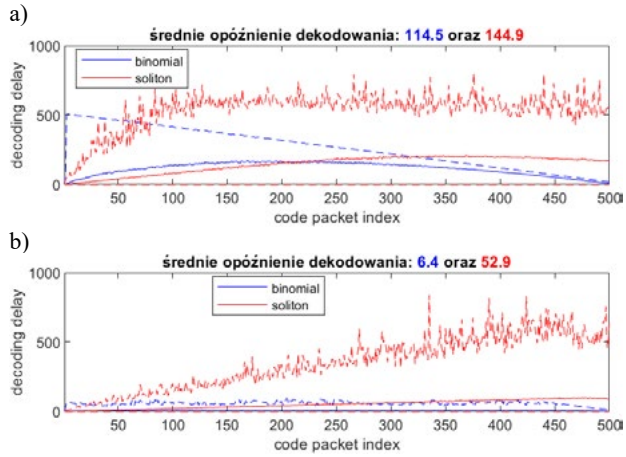
Sytuacja przedstawia się znacznie mniej korzystnie dla rozkładu solitonowego (rys. 7), który cechują mniejsze wagi kolumn. W tym przypadku, po otrzymaniu 500 pakietów kodowych zdekodowanych zostało blisko 300 pakietów, czyli więcej niż dla rozkładu dwumianowego. Jednak tempo dekodowania nowych pakietów utrzymuje się nadal na tym samym poziomie by pod koniec jeszcze znacznie zwolnić. Wszystkie dane nadane zostają zdekodowane dopiero po odebraniu maksymalnie 1241 pakietów, gdzie dla rozkładu dwumianowego cała wiadomość zostaje zdekodowana po otrzymaniu 521 pakietów kodowych (tab. 2).

Rys. 7. Liczba zdekodowanych pakietów dla rozkładu solitonowego i pakietów napływających do kodera stopniowo

Tab.2. Liczba przesłanych pakietów kodowych wymaganych dla osiągnięcia wybranych wartości P_d

P_d	t_0	rozkład dwumianowy	rozkład solitonowy
1,0	0	521	1241
	1		1203
	2		1177
	3		1125
	4		1045
0,95	0	512	1021
	1		954
	2		913
	3		894
	4		878

Bardzo ciekawe zjawisko można zaobserwować, jeżeli przyjmiemy, że pakiety kodowe generujemy szybciej niż napływają pakiety źródłowe. Zauważmy, że opóźnienie dekodowanie dla poprzedniego przypadku (rys. 8a) cechują duże wartości średnie oraz maksymalne, malejące dla rozkładu dwumianowego dla wzrastającego numeru pakietu źródłowego. Jednak, gdy dla rozkładu dwumianowego zwiększymy stosunek liczby pakietów kodowych do liczby pakietów źródłowych do $\alpha = 6/5$ (rys. 8b) opóźnienie dekodowania poszczególnych pakietów gwałtownie spada i pakiety źródłowe są dekodowane na bieżąco z niewielkim opóźnieniem, co umożliwia na zrealizowanie transmisji strumieniowej bez założonej z góry liczby pakietów źródłowych i jednocześnie niedużej złożoności dekodowania. Rozkład solitonowy nawet dla $\alpha = 10$ (rys. 8b) nie pozwala na uzyskanie takiego efektu.



Rys. 8. Opóźnienie dekodowania dla (a) $\alpha = 1$ oraz (b) $\alpha = 6/5$ (rozkład dwumianowy) oraz $\alpha = 10$ (rozkład solitonowy)

4. PODSUMOWANIE

W pracy przedstawiono wyniki badania właściwości kodu fontannowego dla dwóch różnych rozkładów wag kolumn losowej macierzy generującej. Wykazano, że rozkład solitonowy o niskich wagach kolumn, chociaż zmniejsza złożoność dekodowania skutkuje znacznym zwiększeniem opóźnienia zdekodowania całej wiadomości źródłowej. Jednocześnie rozkład dwumianowy o większej średniej wadze kolumn umożliwia uzyskanie efektu transmisji strumieniowej, w której kolejne pakiety źródłowe napływają do kodera stopniowo, o małym opóźnieniu dekodowania bez konieczności stosowania dodatkowego kodera wstępnego.

LITERATURA

[1] Bello Philip. 1963, „Characterization of randomly time-variant channels", IEEE Transactions on Communications Systems.
[2] Haustein Kamila 2021. "Aplikacja demonstrująca działanie kodów fontannowych", projekt dyplomowy inżynierski, Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska.
[3] Ho Tracey, 2003, „Summary of raptor codes", Scientific Commons.
[4] Joshi G. i inni, 2010, „Fountain codes", Global telecommunications conference.
[5] MacKay David JC, 2005, „Fountain codes", IEE Proceedings-Communications, s. 1062-1068.
[6] Nagasubramanian K., 2008: „Code design for erasure channels with limited or noisy feedback", PhD Thesis. Texas A & M University.
[7] Zhang K., Z. Qinyu, J. Jian, 2017, "Bounds on the reliability of RaptorQ codes in the finite-length regime", IEEE Access, 5, 24766-24774.
[8] Chen Guo Tai, 2012, "Analysis of robust soliton distribution for LT code", IEEE 11th International Conference on Signal Processing, vol. 2, pp. 1546-1549.

WYZNACZENIE WARTOŚCI MOCY SYGNAŁU MASKUJĄCEGO, CHRONIĄCEGO SIEĆ PRZED NAMIERZENIEM PRZY WYKORZYSTANIU ALGORYTMU MUSIC

DETERMINING THE POWER VALUES OF A MASKING SIGNAL, PROTECTING THE NETWORK FROM TRACKING USING THE MUSIC ALGORITHM

Jaromir Gajewski[1]; Jerzy Łopatka[2];

[1] Wojskowa Akademia Techniczna, Warszawa, jaromir.gajewski@wat.edu.pl
[2] Wojskowa Akademia Techniczna, Warszawa, jerzy.lopatka@wat.edu.pl

DOI: 10.15199/59.2023.4.84

Streszczenie: **W artykule przedstawiono wyznaczone wartości mocy sygnału uniemożliwiającego wyznaczenie lokalizacji węzłów sieci radiowej przy pomocy algorytmu MUSIC. Omówiono sposób przeprowadzania pomiarów mocy sygnału maskującego oraz przedstawiono otrzymane wyniki. Wszystkie przedstawione symulacje i obliczenia wykonano w środowisku MATLAB.**

Abstract: **The paper presents the determined values of the power of a signal that prevents the localization of radio network nodes using the MUSIC algorithm. The method of measuring the power of the masking signal was discussed, and the obtained results were presented. All presented simulations and calculations were performed in the MATLAB environment.**

Słowa kluczowe: **algorytm MUSIC, wyznaczanie kierunku, walka elektroniczna, przeciwdziałanie elektroniczne.**

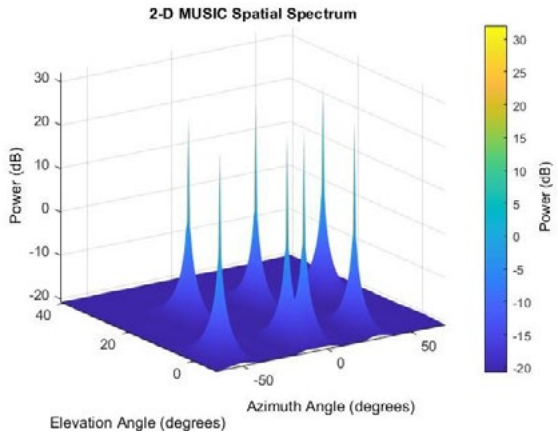
Keywords: **MUSIC algorithm, direction finding, electronic warfare, electronic counteraction.**

1. WSTĘP

W systemach funkcjonujących w dynamicznie zmieniających się warunkach, łączność bezprzewodowa jest jedynym możliwym do zastosowania środkiem komunikacyjnym. W szczególności tyczy się to sieci systemów wojskowych, gdzie poprzez odpowiednią zmianę sposobu funkcjonowania system musi dostosowywać się do takich czynników jak zmiana środowiska propagacyjnego, zmienna konfiguracja systemu połączeń czy wysoka mobilność użytkowników systemu.

Ze względu na konieczność wykorzystania łączności bezprzewodowej w tak szerokim spektrum zastosowań, opracowanych zostało wiele metod rozpoznania takich sieci poprzez ich nasłuchiwanie, dozorowanie, bądź namierzanie . Jednym z najczęściej wykorzystywanych obecnie narzędzi do estymacji kierunku przychodzenia sygnałów (DOA – Direction of Arrival) w środowisku wielodrożnym jest algorytm MUSIC (MUltiple SIgnal Classification) [1,4,5,6], który umożliwia jednoczesne wyznaczenie azymutów wielu sygnałów znajdujących się na tej samej częstotliwości. Algorytm MUSIC bazuje na rozłożeniu macierzy kowariancji sygnału na podprzestrzenie sygnałową i szumową. Główna idea polega na założeniu, że kierunki przychodzenia sygnałów są ortogonalne w stosunku do podprzestrzeni szumowej. W związku z tym funkcja pseudospektrum MUSIC osiąga swoje maksima dla rzeczywistych DOA odpowiadających poszczególnym sygnałom.

Algorytm rozpoczyna pracę od zebrania próbek sygnału z zestawu anten (lub innych sensorów). Następnie tworzona jest macierz kowariancji dla tych próbek. Macierz kowariancji reprezentuje wzajemne zależności między różnymi sygnałami. Jest ona następnie dekomponowana na podprzestrzenie sygnałową i szumową, zazwyczaj za pomocą dekompozycji wartości własnej (Eigen decomposition). Na podstawie otrzymanych podprzestrzeni tworzona jest funkcja pseudospektrum. Podczas wyznaczania jej dla różnych kierunków osiąga ona swoje maksima dla wartości DOA poszczególnych sygnałów. W końcu algorytm MUSIC wykorzystuje funkcję pseudospektrum do estymacji kierunków przychodzenia sygnałów, co można zaobserowować na poniższej charakterystyce (Rys. 1), gdzie osie poziome opisują kąt elewacji i azymutu, natomiast wartości szczytowe pseudowidma MUSIC odpowiadają DOA poszczególnych sygnałów docierających do szyku antenowego odbiornika.



Rys. 1 - przykładowe zobrazowanie pseudowidma MUSIC

Algorytm MUSIC ma wiele zastosowań w wielu dziedzinach, a ze względu na jego zdolność do rozróżniania i śledzenia źródeł sygnału w trudnych warunkach jest on podstawą wielu nowoczesnych technologii pomimo swojej złożoności.

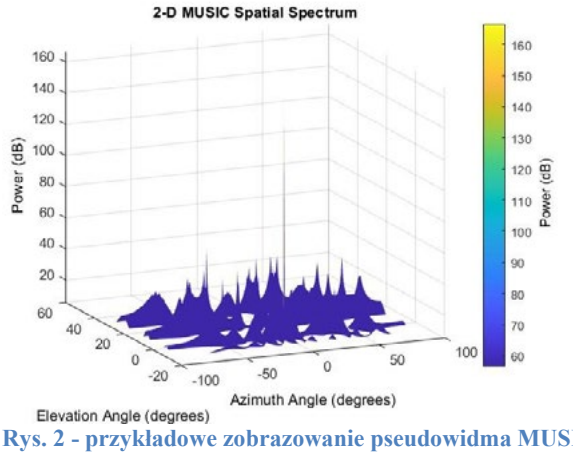
W niniejszym artykule przedstawiono ideę maskowania sieci radiowej przy pomocy sygnału maskującego, scenariusz badania mającego na celu wyznaczenie wartości tej mocy oraz otrzymane wyniki.

2. SYGNAŁ MASKUJĄCY

2.1. Idea sygnału maskującego

W wielu systemach, w szczególności w systemach wojskowych koniecznym bywa ukrycie węzłów pracującej sieci radiowej przed środkami namierzania radiowego przeciwnika. Można to uzyskać poprzez przerwanie pracy sieci radiowej, co jednak uniemożliwia wykonanie wielu, często niezbędnych zadań.

Ideą zastosowania sygnału maskującego jest wprowadzenie sygnału o określonych parametrach i wystarczającej mocy aby uniemożliwić rozpoznanie azymutów pracujących węzłów sieci. Takie rozwiązanie może umożliwić ukrycie pracującej sieci radiowej bez konieczności przerywania jej pracy co może się okazać kluczowe w wielu sytuacjach występujących na polu walki. Opisane zjawisko ilustruje rysunek 2.



Rys. 2 - przykładowe zobrazowanie pseudowidma MUSIC podczas pracy urządzenia emitującego sygnał maskujący

2.2. Założenia pomiaru mocy sygnału maskującego

Badania przeprowadzono w środowisku MATLAB na symulowanej sieci pracującej na częstotliwości 400MHz, złożonej z 6 nieruchomych węzłów, których maksymalna odległość kątowa względem urządzenia namierzającego wynosiła 100º, a sygnały emitowane przez węzły sieci są sygnałami nieskorelowanymi.

Jako urządzenie namierzające wykorzystano zamodelowany jednolity prostokątny szyk anten o wymiarach 11x11, której elementy rozmieszczono w równomiernych odstępach wzdłuż dwóch osi, równych połowie długości fali częstotliwości nośnej sygnału emitowanego w symulowanej sieci. Wykorzystanie szyku prostokątnego umożliwia metodzie MUSIC określenie nie tylko azymutów węzłów pracujących w sieci radiowej, ale również kątów elewacji pod jakimi padają one na szyk antenowy[2].

Do symulacji sygnałów odebranych przez szyk antenowy wykorzystana została funkcja collectPlaneWave, która oblicza odpowiedzi elementów szyku antenowego na jedną lub więcej fal płaskich, które docierają do tablicy pod określonymi kątami. Fale płaskie są modelowane jako sygnały sinusoidalne o określonej częstotliwości, które docierają do szyku antenowego, a funkcja uwzględnia różnice czasowe pomiędzy sygnałami docierającymi do różnych elementów szyku.

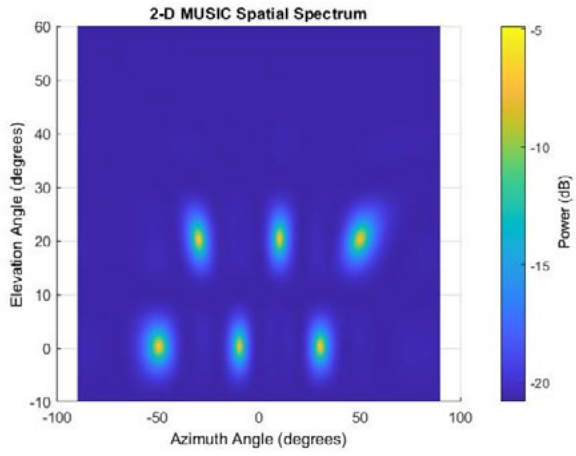
Do wykonania 2-wymiarowej estymacji kierunku przychodzenia sygnałów (DOA) przy użyciu algorytmu MUSIC wykorzystano funkcję phased.MUSICEstimator2D, która tworzy obiekt implementujący algorytm MUSIC. Ilość namierzanych sygnałów nie została podana w niniejszej funkcji, gdyż w rzeczywistych warunkach urządzenie namierzające nie posiada wiedzy o ilości urządzeń w sieci. Z tego powodu badania przeprowadzono dla dwóch algorytmów określania ilości sygnałów:

- MDL (Minimum Description Length)[3,4], algorytm minimalnej długości opisu jest kryterium informacyjnym służącym do wyboru modelu statystycznego, które minimalizuje sumę opisu danych modelu. Dla każdej możliwej liczby sygnałów algorytm oblicza koszt opisu danych i modelu. Model z najniższym kosztem jest wybierany jako pierwszy.

- AIC (Akaike Information Criterion)[3,4], jest metodą stosowaną do określania liczby sygnałów w estymacji MUSIC. Dla każdej możliwej liczby sygnałów algorytm AIC oblicza koszt związany z błędem estymacji i złożonością modelu. Model z najniższym kosztem jest wybierany jako pierwszy.

3. SCENARIUSZ BADANIA

Badania rozpoczęto od wyznaczenia minimalnej wartości SNR dla których algorytm MUSIC jest w stanie oddzielić przestrzeń sygnałową od szumowej. Wyznaczono, że jest to wartość SNR = -11 dB. Węzły sieci umieszczono względem urządzenia namierzającego w następujących odległościach kątowych: (-10;0), (10;20), (-30;20), (30;0), (-50;0), (50;20), gdzie wartości te określają odpowiednio (azymut;elewacja). Poniższe psudowidmo MUSIC ilustruje położenie węzłów opisanych powyżej (Rys. 3).



Rys. 3 - położenie węzłów symulowanej sieci radiowej

Następnie wprowadzono dodatkowe źródło sygnałowe umieszczone pomiędzy źródłami pracującej sieci (0;0) takie, aby jego sygnał był nieskorelowany z sygnałami emitowanymi przez symulowaną sieć radiową.

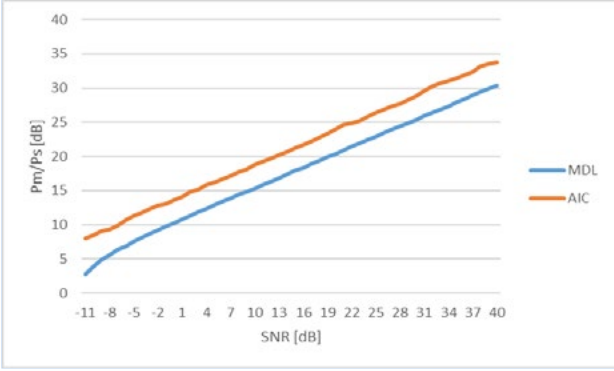
W kolejnym kroku zwiększano moc sygnału maskującego tak długo, aż wyniki działania algorytmu MUSIC uniemożliwiały uzyskanie rzeczywistych wyników. Ze względu na losowy charakter szumu obecnego w środowisku propagacyjnym niemożliwe jest wyznaczenie jednej wartości mocy sygnału maskującego, dlatego w niniejszej pracy za kryterium przyjęto mniejszą niż 50% skuteczność otrzymania pseudowidma algorytmu MUSIC umożliwiającego zlokalizowanie któregokolwiek z węzłów sieci na próbie 100 pomiarów.

Pomiary przeprowadzono dla wartości SNR w zakresie (-11,40) dB.

4. WYNIKI BADANIA

4.1. Otrzymane wyniki

Wyniki pomiarów przedstawiono w postaci stosunku mocy sygnału maskującego (*Pm*) do mocy sygnałów emitowanych w symulowanej sieci (*Ps*)w funkcji wartości SNR, dla których przeprowadzane były pomiary.



Rys. 4 - wyznaczone wartości mocy sygnału maskującego przedstawione w postaci stosunków mocy sygnału maskującego (Pm) do mocy sygnałów emitowanych w sieci radiowej (Ps) dB w funkcji SNR dB

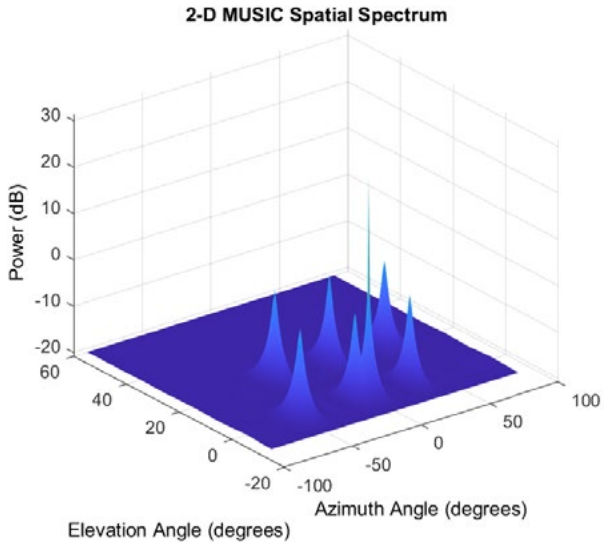
Rysunek 4 przedstawia charakterystykę, wymagana moc niezbędna do zamaskowania sieci przed urządzeniem namierzającym działającym w oparciu o realizacją algorytmu MUSIC rośnie wraz ze wzrostem wartości współczynnika SNR. Widać również, że namiernik działający w oparciu o algorytm określania ilości źródeł promieniowania AIC okazuje się skuteczniejszy w namierzaniu sygnałów maskowanych sygnałem maskującym. Średnia różnica pomiędzy wymaganym stosunkiem mocy sygnału maskującego do mocy sygnałów w sieci radiowej dla dwóch badanych algorytmów wynosi 3,54 dB

Warto zauważyć, że dla niskich SNR metoda MUSIC przestaje dawać oczekiwane wyniki już przy niewielkich poziomach mocy sygnału maskującego.

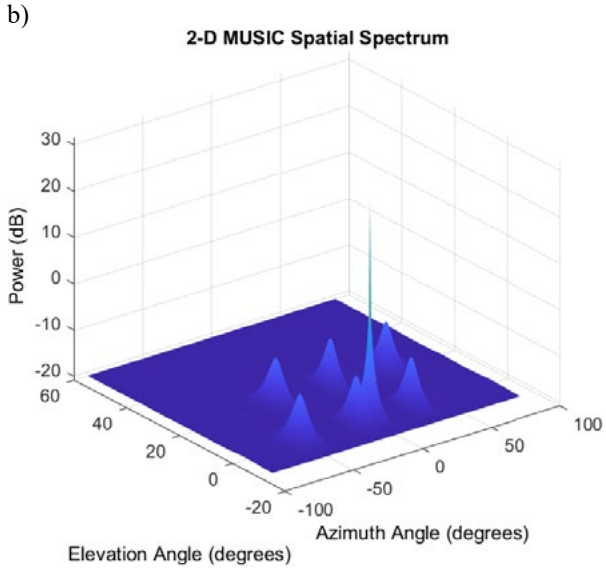
4.2. Zachowanie pseudowidma MUSIC dla zwiększającej się wartości mocy sygnału maskującego.

Wraz ze wzrostem wartości mocy sygnału maskującego lokalne maksima znajdujące się w miejscach pozostałych źródeł wypłaszczają się i stają się coraz trudniejsze do dostrzeżenia. Opisane zjawisko zilustrowane jest na poniższych charakterystykach, gdzie dla różnych wartości sygnału maskującego przedstawiono charakterystykę pseudospektrum MUSIC. Charakterystyki wyskalowano tak, aby wskazywały pseudowidmo MUSIC od poziomu -20 dB, natomiast wartości sygnału maskującego przedstawiono w postaci stosunku jego mocy do mocy sygnałów emitowanych w sieci

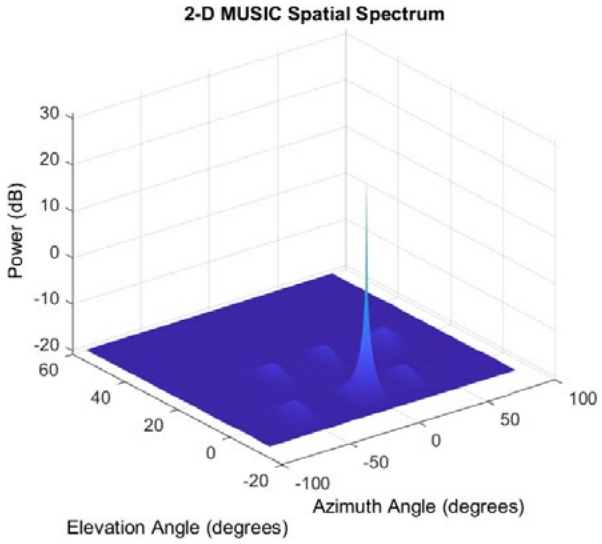
a)



b)



c)



Rys. 5 - zobrazowania pseudowidma MUSIC dla rosnących wartości sygnału maskującego przy stałym współczynniku SNR = 11 dB a) 13 dB b) 16 dB c) 18 dB

5. WNIOSKI

Wprowadzenie sygnału maskującego może skutecznie zakłócić pracę namiernika radiowego działającego w oparciu o algorytm MUSIC. Wyniki należałoby jednak zweryfikować poprzez przeprowadzenie pomiarów rzeczywistych i odniesieniu ich do wyników uzyskanych w symulacji.

Wraz ze wzrostem wartości mocy sygnału maskującego lokalne maksima pseudospektrum MUSIC wypłaszczają się stając się coraz mniej widocznymi na tle sygnału maskującego.

Można zaobserwować, że wykorzystanie metody określania liczby źródeł sygnałów AIC było efektywniejsze niż w przypadku metody MDL, a średnia różnica dla tych metod wynosi 3,54 dB.

Wyniki symulacji pokazują, że dla wartości SNR = -11 dB moc sygnału maskującego umożliwiającego skuteczną ochronę pracującej sieci radiowej dla algorytmu AIC wynosi 8 dB w porównaniu z mocą sygnałów emitowanych w sieci, natomiast dla algorytmu MDL 2,8 dB. Wartość ta rośnie liniowo wraz ze wzrostem współczynnika SNR i dla jego wartości równej 40 dB dla algorytmu AIC wynosi 33,7 dB, a dla MDL 30,4 dB.

Wyznaczone wartości mocy sygnału maskującego odnoszą się do źródła umieszczonego w bezpośredniej bliskości pracującej sieci. Umieszczenie źródła tego sygnału pomiędzy siecią, a urządzeniem namierzającym np. na platformie UAV może znacząco zmniejszyć wymagany do maskowania poziom sygnału, tym bardziej im bliżej urządzenia namierzającego by się ona znajdowała.

LITERATURA

[1] Bhavani, R., & Ramanujam, R. (2010). A Robust MUSIC Algorithm Against Impulsive Noise and Subspace Swapping. In 2010 Second International Conference on Computing, Communication and Networking Technologies(pp.1-5). IEEE

[2] Schmidt, R. (1986). Multiple emiter location and signal parameter estimation. IEEE transactions on antennas and propagation, 34(3),276-280

[3] Stoica, P., & Moses, R. L. (2005). Spectral Analysis of Signals. Prentice Hall.

[4] Wax, M., & Ziskind, I. (1988). On unique localization of multiple sources by passive sensor arrays. IEEE Transactions on Acoustics, Speech, and Signal Processing, 36(7),996-1000.

[5] Qu Zhiyu, Li Mingshi (2022) A Polarization MUSIC Algorithm Based on Cylindrical Conformal Arrays

[6] Khan Nabeel Ali, Ali Sadiq, Mohammadi Mokhtar, Haneef Muhammad (2020) Direction of arrival estimation of sources with intersecting signature in time-frequency domain using a combination of IF estimation and MUSIC algorithm

SENSOR MONITORINGU WIDMA ZINTEGROWANY Z BEZZAŁOGOWYM STATKIEM POWIETRZNYM

SPECTRUM MONITORING SENSOR INTEGRATED WITH AN UNMANNED AERIAL VEHICLE

Maciej Mazuro[1],

[1]Wojskowa Akademia Techniczna, Wydział Elektroniki, Warszawa, maciej.mazuro@wat.edu.pl

DOI: 10.15199/59.2023.4.85

Streszczenie: **W ramach II edycji konkursu o nagrodę Ministra Obrony Narodowej (MON) na realizację Bezzałogowego Systemu Powietrznego (BSP) do zastosowań związanych z obronnością i bezpieczeństwem państwa opracowałem sensor monitoringu widma w czasie rzeczywistym na platformie BSP pk. System MARS, który jest demonstratorem technologii sensingu w zakresie częstotliwości od 70 do 6000 MHz. Rozwiązanie opiera się o integrację radia definiowanego programowo z mikrokomputerem Raspberry Pi oraz umieszczeniem na bezzałogowej platformie latającej.**

Abstract: **As part of the second edition of the competition for the award of the Minister of National Defense for the implementation of the Unmanned Aerial System (UAV) for applications related to national defense and security, I developed a real-time spectrum monitoring sensor on the UAV platform under the codename MARS System, which is a technology demonstrator sensing in the frequency range from 70 to 6000 MHz. The solution is based on the integration of a software-defined radio with Raspberry Pi microcomputer and placement on an unmanned aerial platform.**

Słowa kluczowe: **bezzałogowy system powietrzny, GNU Radio Companion, monitoring widma, radio definiowane programowo, rozpoznanie radioelektroniczne.**

Keywords: **GNU Radio Companion, signal intelligence, software-defined radio, spectrum sensing, unmanned aerial system.**

1. WSTĘP

Kolejna edycja konkursu MON pozwoliła młodym inżynierom sprawdzić swoje umiejętności i zaprojektować systemy mające zastosowania w obronności i bezpieczeństwie państwa. W proponowanym rozwiązaniu podjąłem się implementacji algorytmu monitoringu widma w mikrokomputerze Raspberry Pi z wykorzystaniem radia programowalnego USRP B200 mini. Sensor został umieszczony na komercyjnej platformie bezzałogowej – DJI Mavic 3 Cine Premium Combo. Detektor energii, dzięki integracji mikrokomputera z odbiornikiem radiowym realizował sensing w czasie rzeczywistym w zakresie częstotliwości od 70 do 6000 MHz. Możliwości techniczne statku powietrznego ograniczały pracę systemu pk. System MARS (Rys. 1) do lotów trwających nie dłużej niż 30 minut.

Rys. 1. Logo systemu pk. MARS

Monitoring spektrum elektromagnetycznego może opierać się o wiele algorytmów [3], ale ze względu na wcześniejsze doświadczenie i implementację rozwiązania w swojej pracy magisterskiej [1] wybrałem detektor energii z estymacją poziomu szumu ED-ENP (*ang. Energy Detector with Estimated Noise Power*). Głównymi zaletami detektora ED-ENP jest brak konieczności posiadania wiedzy *a priori* na temat badanego sygnału, niskie wymagania obliczeniowe oraz mała złożoność implementacyjna [2]. Do ewaluacji skuteczności zaimplementowanego rozwiązania zostały wykorzystane charakterystyki operacyjne odbiornika ROC (*ang. Receiver Operation Characteristic*) oraz wykresy zależności prawdopodobieństwa detekcji oraz fałszywego alarmu w funkcji SNR (*ang. Signal to Noise Ratio*).

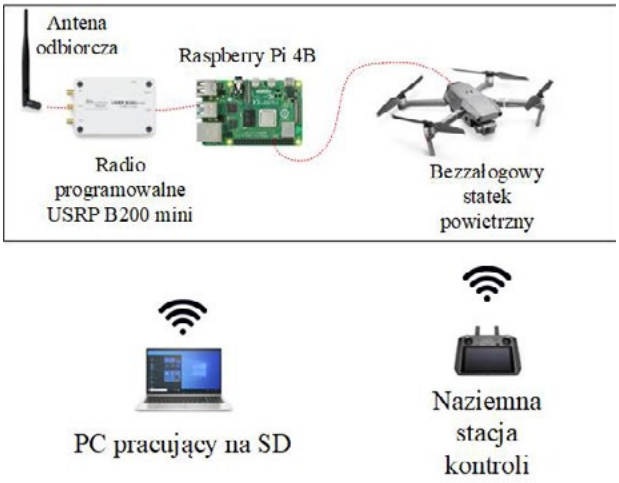
Systemy został oparty o rozwiązania komercyjne typu COTS (*ang. Commercial Off The Shelf*), czyli podzespoły dostępne do użytku prywatnego. Zaletą takiego podejścia jest łatwy dostęp do technologii i relatywnie niskie ceny poszczególnych elementów. Integrację ułatwił fakt korzystania z oprogramowania GNU Radio Companion, należącego do klasy *OpenSource*. Pozwoliło to na modyfikację kodu źródłowego zgodnie z wymaganiami i możliwościami sprzętowymi. Do analizy efektów sensingu i zobrazowania wyników wykorzystałem oprogramowanie Matlab.

Przeprowadzone testy i otrzymane wyniki potwierdziły skuteczność proponowanego rozwiązania i zostaną omówione w dalszej części artykułu. Ze względu na charakter demonstracyjny system obarczony jest ograniczeniami – wybrany bezzałogowy statek powietrzny, po obciążaniu go dodatkowymi elementami może realizować zadania w ograniczonym czasie, nie przekraczającym 30 minut. Kolejnym istotnym czynnikiem podczas planowania i stawiania zadań jest zasięg działania – łączność radiowa ze statkiem powietrznym jest ograniczona zasięgiem sieci Wi-Fi.

2. ZAŁOŻENIA PROJEKTOWE

2.1. Koncepcja systemu

System został zbudowany zgodnie ze schematem (Rys. 2), który przedstawia połączenia pomiędzy poszczególnymi elementami. Komputer wraz z kontrolerem ruchu tworzy stanowisko dowodzenia (SD), które kontroluje i zarządza pracą całego systemu. Mikrokomputer Raspberry Pi pracuje w jednej sieci z naziemną stacją i przesyła odebrane próbki sygnałów w czasie rzeczywistym dzięki wykorzystaniu protokołu TCP/IP oraz połączenia klient-serwer. Otrzymane dane są przetwarzane przy wykorzystaniu oprogramowania GNU Radio Companion oraz skryptu w języku Python. Naziemna stacja kontroli odpowiada za sterowanie lotem BSP.



Rys. 2. Schemat systemu

Program pozwala na zobrazowanie kilku rodzajów przebiegów – w dziedzinie czasu, częstotliwości (spektrogram), wodospad (*ang. waterfall*) oraz wykresy konstelacji próbek IQ (*ang. In-phase and Quadrature*). Dodatkowy skrypt zapisuje otrzymane wyniki prawdopodobieństwa detekcji w zależności od oszacowanego poziomu SNR. Służą one do dalszej ewaluacji oraz zestawienia rezultatów otrzymanych testów.

2.2. Detektor energii ED-ENP

Detektor energii z estymacją poziomu szumu jest skuteczny przy założeniu, że odbierane zakłócenia mają charakter addytywnego białego szumu gaussowskiego. W tym przypadku odbierany sygnał można zapisać jako:

$$y(n) = s(n) + w(n), \tag{1}$$

gdzie $s(n)$ to próbki odbieranego sygnału, a $w(n)$ to próbki szumowe.

Statystyka testowa na bazie energii odbieranego sygnału jest określona jako:

$$M = \sum_{n=0}^{N} |y(n)|^2, \tag{2}$$

gdzie N to ilość próbek. Jeżeli otrzymana wartość przekroczy określony próg detekcji λ stwierdza się detekcję sygnału. Prawdopodobieństwo detekcji określa prawdopodobieństwo stwierdzenia detekcji sygnału kiedy jest on w rzeczywistości obecny, a prawdopodobieństwo fałszywego alarmu określa prawdopodobieństwo stwierdzenia o większej wartości statystyki testowej od wartości progowej, gdy w rzeczywistości jest odwrotnie. Prawdopodobieństwo detekcji P_d określone jest wzorem:

$$P_{d,ED-ENP} = Q\left(\frac{\frac{\lambda}{1+SNR}-1}{\sqrt{\frac{N+M}{N \cdot M}}}\right), \tag{3}$$

gdzie M to liczba próbek wykorzystanych do oszacowania poziomu szumu, a Q to funkcja Gaussa opisana zależnością: $Q(x) = 1/2erfc(x/\sqrt{2})$. Estymacja wartości SNR następuje poprzez wyliczenie średniej energii odebranego sygnału i pomniejszenie jej o wartość wariancji sygnału. Jest to powszechnie stosowana metoda do oszacowania wartości zakłóceń bez znajomości warunków propagacyjnych w kanale radiowym.

Tabela (Tab. 1) zawiera przykładowe rezultaty detekcji z informacją na temat częstotliwości środkowej, na której prowadzony był sensing, wartości SNR oraz otrzymanego P_d. Wartość P_{fa} dla poniższych wyników wynosiła 0.1. Pozwala to zobrazować jak zmieniały się wartości prawdopodobieństwa detekcji w zależności od poziomu zakłóceń w kanale.

Tab. 1. Wyniki detekcji dla $P_{fa} = 0.1$

Częstotliwość	*SNR*	P_d
102,5MHz	0 dB	1.00
102,5MHz	2 dB	1.00
102,5MHz	-1 dB	0.98
102,5MHz	-2 dB	0.97
102,5MHz	-3 dB	0.95
102,5MHz	-4 dB	0.88

Szczegółowa analiza otrzymanych wyników wraz z otrzymanymi charakterystykami zostanie przeprowadzona w kolejnym rozdziale.

3. OCENA EFEKTYWNOŚCI

3.1. Metodyka oceny skuteczności detekcji

Podstawowym narzędziem oceny skuteczności zaimplementowanego detektora energii są charakterystyki operacyjne ROC, które przedstawiają zależności pomiędzy prawdopodobieństwem detekcji P_{d}, prawdopodobieństwem fałszywego alarmu P_{fa} oraz wartościami SNR. Teoretyczne charakterystyki ROC [1], otrzymane na podstawie symulacji ukazują pracę odbiornika w szerokim zakresie wartości stosunku sygnału użytecznego do sygnału zakłóceń – od -20dB do 20dB.

W czasie pracy z rzeczywistymi sygnałami radiowymi, odbieranymi przez radio programowalne USRP B200 mini w przeważającej ilości przypadków sygnał użyteczny miał większą moc niż zakłócenia. Z tego powodu konieczne było manipulowanie wartościami mocy zakłóceń, aby uzyskać ujemne wartości SNR. Po ustaleniu oczekiwanej wartości P_{fa} oraz przestrojeniu odbiornika USRP na żądaną częstotliwość wartość mocy szumu była zwiększana w kodzie programu. W ten sposób uzyskiwałem detekcję sygnału przy niższych wartościach SNR, co w prawdziwych warunkach było utrudnione. Gdy w kanale radiowym nie pojawiał się sygnał użyteczny poziom energii odbieranego sygnału nie przekraczał wartości progowej i nie następowała detekcja sygnału.

Zebrane w ten sposób wyniki P_{d} zapisywane były w bazie danych, a następnie analizowane przy użyciu środowiska programistycznego Matlab. Analizie poddane zostały trzy wartości P_{fa}: 0.1, 0.05 oraz 0.02. Testy skuteczności sensora prowadzone były w trakcie lotu BSP na terenie Starego Bemowa i okolicznych miejscowości. Analizie poddane zostały kanały radiowe o następujących częstotliwościach:

- 100MHz-110MHz – systemy rozgłoszeniowe radiofonii

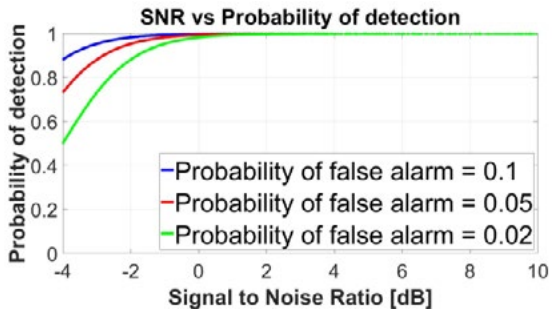
- 460MHz-462MHz – system trunkingowy łączności służb wewnętrznych

- 790MHz-820MHz – stacje bazowe operatorów komórkowych

- 890MHz – 915MHz – system GSM dla łącza w górę (up-link).

3.2. Uzyskane rezultaty

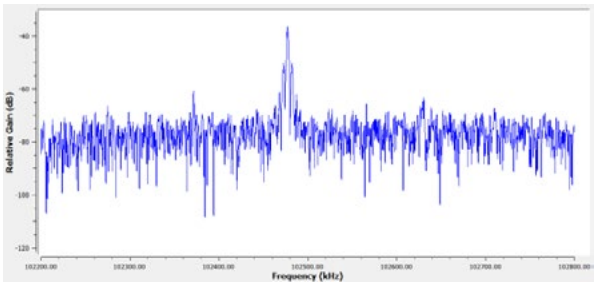
Opracowany sensor realizuje monitoring widma na podstawie detektora energii z estymacją poziomu szumu z zadowalającymi rezultatami. Na bazie przeprowadzonych testów i otrzymanych wyników stwierdzam zgodność z założeniami teoretycznymi. Realizacja zadań w czasie rzeczywistym pozwala na otrzymanie dużych zbiorów danych, ale w artykule przedstawione zostaną tylko ich część. Rysunek (Rys. 3) przedstawia charakterystyki operacyjne odbiornika uzyskane podczas pomiarów dla kanału radiowego o częstotliwości środkowej wynoszącej 101,5 MHz i dla różnych wartości P_{fa} oraz SNR.



Rys. 3. Charakterystyka pracy odbiornika

Na powyższym zobrazowaniu widać różnicę otrzymywanych wartości P_{d} w funkcji zakładanego P_{fa} oraz zmieniających się warunków kanałowych i wartości SNR. W ramach przeprowadzonych testów widoczna jest zależność prawdopodobieństwa detekcji od prawdopodobieństwa fałszywego alarmu, co potwierdza teoretyczne rozważania. Zmniejszenie P_{fa} , a co za tym idzie zwiększenie wymagań wobec systemu skutkuje spadkiem wartości P_{d}. Dobrze widoczne jest to dla $SNR = -2dB$, gdzie sensor z $P_{\text{fa}} = 0.1$ otrzymuje $P_{\text{d}} = 0.97$, w przypadku $P_{\text{fa}} = 0.05$ wartość P_{d} obniża się do 0.95, a dalsze obniżanie P_{fa} do wartości 0.02 skutkuje otrzymaniem $P_{\text{d}} = 0.88$. Dla mniejszych wartości SNR tendencja spadku wartości P_{d} wraz z obniżeniem wartości P_{fa} jest jeszcze bardziej widoczna. W przypadku systemu pracującego w warunkach kanałowych odpowiadającym $SNR = -4dB$ otrzymujemy następujące pary wyników: $P_{fa} = 0.1\ i\ P_{\text{d}} = 0.88$, $P_{fa} = 0.05\ i\ P_{\text{d}} = 0.78$ oraz $P_{fa} = 0.02\ i\ P_{\text{d}} = 0.5$. Spadek wartości stosunku mocy sygnału użytecznego do mocy zakłóceń o 2dB skutkuje zmniejszeniem prawdopodobieństwa detekcji o średnia 21 punktów procentowych. Wynik ten znacząco podwyższa sensor pracujący z $P_{fa} = 0.02$, w którym spadek P_{d} nastąpił z wartości 0.88, aż do 0.5. W przypadku wzrostu wartości SNR zmiana jest zauważalna tylko do pewnego stopnia. Dla $SNR = 0dB$ tylko sensor o największych wymaganiach nie otrzymuje 100% prawdopodobieństwa detekcji. Wraz z dalszym zwiększaniem wartości stosunku mocy sygnału użytecznego do zakłóceń prawdopodobieństwo detekcji dąży do jedności i osiąga ją we wszystkich przypadkach.

Skuteczność działania zaproponowanego sensora była potwierdzona również w trakcie działania w czasie rzeczywistym przez implementacje zobrazowań w dziedzinie częstotliwości w środowisku programistycznym GNU Radio Companion. Spektrogram na podstawie otrzymanych wyników przedstawiony jest poniżej (Rys. 4).



Rys. 4. Otrzymane charakterystyki w GNU Radio

Uzyskane spektrogramy uwidaczniały cechy charakterystyczne próbek odbieranych sygnałów, a w przypadku stwierdzenie detekcji sygnału na zadanej częstotliwości na przebiegach pojawiały się charakterystyczne ‘prążki’, które pokrywały się z wynikami otrzymanymi w detektorze energii ED-ENP.

4. WNIOSKI

Proponowany system pk. MARS stanowi nowatorskie rozwiązanie w zakresie realizacji zadań rozpoznania radioelektronicznego w integracji z bezzałogowymi statkami powietrznymi. Obecnie na wyposażeniu Sił Zbrojnych Rzeczypospolitej Polskiej większość systemów SIGINT (*ang. Signal Intelligence*) jest stacjonarna bądź zabudowana na bazie pojazdów kołowych lub opancerzonych. W nowoczesnych armiach państw zachodnich widoczna jest tendencja przenoszenia systemów rozpoznawczych na platformy bezzałogowe, zarówno lądowe, morskie jak i powietrzne. Podbudowane jest to chęcią minimalizacji strat ludzkich i sprzętowych. Śmigłowiec, odrzutowiec, samolot transportowy to zakupy wielomilionowe. Bezzałogowy statek powietrzny, odpowiednio wyposażony i skonfigurowany do pracy może stanowić równie rzetelne źródło informacji.

Wyniki uzyskane przez sensor świadczą o wysokiej skuteczności detekcji, ale nie powinno traktować się ich jako produkt finalny. Należy rozważyć implementacje dodatkowych algorytmów analizy sygnałów, które umożliwią nie tylko stwierdzenie występowania emisji, ale również ich interpretacje. Techniki przechwytywania i dekodowania relacji radiowych są skomplikowane oraz wymagają większych mocy obliczeniowych, stąd w planach rozwoju należy uwzględnić modernizację bądź wymianę pokładowego mikrokomputera na rozwiązanie dedykowane do zadań DSP (*ang. Digital Signal Processing*).

Kolejny aspekt to wykorzystana platforma bezzałogowa. W projekcie zakupiony został dron DJI Mavic 3 Cine Premium Combo, który nie jest przeznaczony do realizacji zadań SIGINT. Integracja radia programowalnego i mikrokomputera Raspberry Pi z platformą stanowiła duże wyzwania i ograniczenie możliwości. Dron nie jest przystosowany do przenoszenia dodatkowego obciążenia. Należy rozważyć zbudowanie nowego statku bezzałogowego według własnego projektu i założeń, który będzie lepiej przystosowany do integracji z dodatkowym wyposażeniem.

W kontekście realizacji sensingu coraz częściej wymienia się kooperacyjny monitoring widma jako rozwiązanie bardziej skuteczne. Akwizycja danych z wielu sensorów pozwala na zwiększenie pewności podejmowanej decyzji. Konieczne do tego jest zaimplementowanie dodatkowych algorytmów fuzji danych z wielu źródeł, co stanowi dodatkowe obciążenie procesorów obliczeniowych pokładowego komputera.

System stanowi demonstrator możliwości prowadzenia rozpoznania radioelektronicznego z pokładu bezzałogowego statku powietrznego, ale powinien być rozwijany i zdobywać nowe możliwości techniczne. Przewaga nad lądowymi bądź stacjonarnymi systemami SIGINT jest znacząca i niepodważalna. Większy zasięg i mobilność oraz lepsze warunki propagacyjne pozwalają na szybsze i efektywniejsze realizowanie zadań.

LITERATURA

[1] Mazuro Maciej. 2022. „Ocena efektywności wykrywania emisji radiowych z wykorzystaniem UAV". *Praca dyplomowa*. Wojskowa Akademia Techniczna.

[2] Skokowski Paweł. 2021. Budowanie świadomości sytuacji elektromagnetycznej w sieciach doraźnych z węzłami kognitywnymi. Wojskowa Akademia Techniczna.

[3] Yucek T., Arslan H. 2009. „A Survey of Spectrum Sensing Algorithms for Cognitive Radio Applications". *IEEE Communications Surveys & Tutorials*, 116–130.

WYZNACZANIE NATĘŻENIA POLA, GĘSTOŚCI MOCY I SAR W OTOCZENIU RADIOKOMUNIKACYJNYCH STACJI BAZOWYCH

DETERMINATION OF FIELD STRENGTH, POWER DENISITY AND SAR IN THE VICINITY OF RADIOCOMMUMICATION BASE STACIONS

Waldemar Karcz

Emitel S.A., Szczecin, waldemar.karcz@emitel.pl

DOI: 10.15199/59.2023.4.86

Streszczenie: **Na podstawie normy IEC 62232 oraz na podstawie rekomendacji ITU zostały przedstawione aspekty dotyczące zgodności poziomów PEM ze stacji bazowych 5G z dopuszczalnymi limitami ICNIRP. Przedstawiono również uproszczone procedury oceny w celu spełnienia wymagań zgodności produktów instalowanych w stacjach bazowych. Testy w komercyjnych i testowych sieciach 5G przeprowadzone przez światowe agencje zdrowia i laboratoria badawcze wykazały, że poziomy PEM z sieci 5G są niskie i znacznie poniżej międzynarodowych limitów bezpieczeństwa.**

Abstract: **Based on the IEC 62232 standard and ITU recommendations, aspects regarding the compliance of EMF levels from 5G base stations with the permissible ICNIRP limits have been presented. Simplified assessment procedures to meet the compliance requirements of products installed in base stations are also presented. Tests in commercial and test 5G networks by global health agencies and research laboratories have shown that EMF levels from 5G networks are low and well below international safety limits**

Słowa kluczowe: **gęstość mocy, limity ekspozycji, małe komórki, natężenie pola**

Keywords: **exposure limits, field strength, power density, small cells**

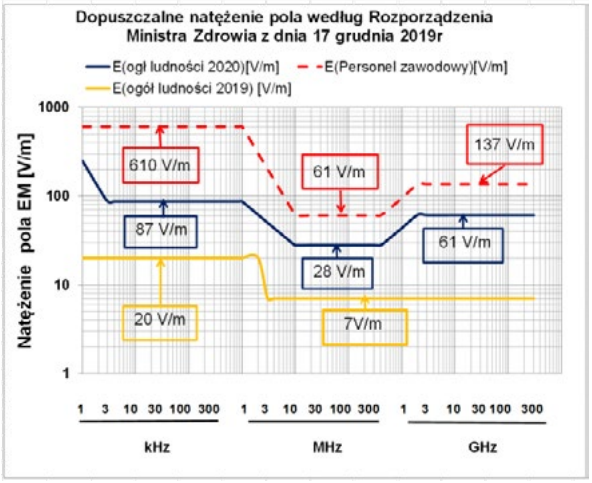
1. WSTĘP

Zmiany dopuszczalnych poziomów pól elektromagnetycznych w środowisku zostały wprowadzone Rozporządzeniem Ministra Zdrowia z dnia 17 grudnia 2019 r. [8] i weszły w życie 1 stycznia 2020 r. Jednocześnie z dniem 1 stycznia 2020 r. przestało obowiązywać Rozporządzenie Ministra Środowiska z dnia 30.10.2003 r. Rozporządzenie to zostało zastąpione Rozporządzeniem Ministra Klimatu z dnia 17 lutego 2020 r. [7] w sprawie sposobów dotrzymywania dopuszczalnych poziomów pól elektromagnetycznych w środowisku. Pozytywna zmiana dopuszczalnych poziomów pól w środowisku wprowadziła nowe limity PEM dla poszczególnych zakresów częstotliwości zgodne z wytycznymi ICNIRP [1]. Niestety pozostałe rozporządzenia nie zostały zaktualizowane i dodatkowo w ramach tak zwanego „sprawnego i szybkiego procesu legislacyjnego" utrzymano w POŚ[10] artykuł 122a oraz utrzymano w mocy obowiązek zgłaszania instalacji radiokomunikacyjnych, których anteny emitują moc EIRP ≤15W. Polska norma PN-EN 62232 [6] zawiera wytyczne zgodne z ICNIRP [1] pozwalające wyznaczać natężenia pola, gęstości mocy oraz SAR w otoczeniu radiokomunikacyjnych stacji bazowych dla oceny poziomu ekspozycji człowieka. Norma Europejska EN 62232:2017 została uznana przez PKN za polską normę PN-EN 62232:2018-01.

2. AKTUALNY STAN PRAWNY

2.1. Dopuszczalne poziomy PEM

Zgodnie z Rozporządzeniem Ministra Zdrowia [8] od 1 stycznia 2020 r. obowiązują nowe dopuszczalne poziomy pól elektromagnetycznych w miejscach dostępnych dla ludności. Dopuszczalne poziomy natężenia pól w miejscach dostępnych dla ogółu ludności przedstawione zostały na Rys. 1. Granica dopuszczalnych poziomów PEM dla personelu zawodowego nie została jeszcze w Polsce wprowadzona.



Rys. 1. Dopuszczalne poziomy PEM w środowisku

2.2. Sposoby dotrzymywania PEM w miejscach dostępnych dla ludności

Artykuł 121 i 122 ustawy POŚ[10] w połączeniu z ust. 5 pkt 2 załącznika do rozporządzenia [7] pozwalają, w sposób zgodny ze stanem faktycznym, zakwalifikować lub wykluczyć daną instalację do mogących negatywnie oddziaływać na środowisko. Postępując zgodnie z art. 3 pkt 21 ustawy POŚ[10] i z ust. 5 pkt 2 załącznika [7] można ustalić czy dana instalacja jest zgodna z dopuszczalnymi limitami. Ochrona przed polami elektromagnetycznymi polega na zapewnieniu jak najlepszego stanu

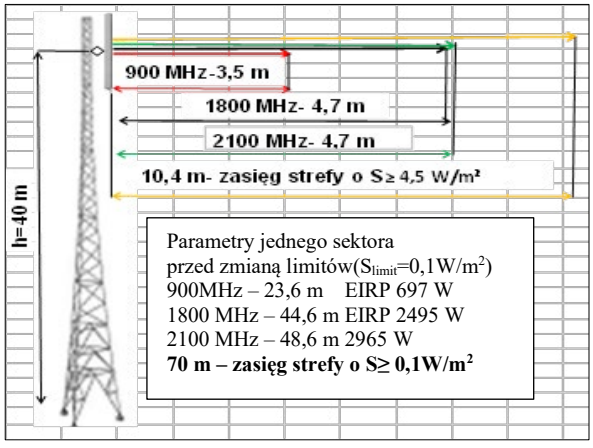
środowiska poprzez utrzymanie poziomów pól elektromagnetycznych poniżej dopuszczalnych limitów. Ten cel osiąga się poprzez:

a) instalowanie anten na wysokich podporach,
b) dobór anten o kształtowanych odpowiednio charakterystykach poziomych i pionowych,
c) dobór optymalnej mocy EIRP,
d) zachowanie odpowiedniej odległości od anten do miejsc dostępnych dla ludności.

Ponadto systematyczny rozwój radiokomunikacji i telekomunikacji oraz rozwój technologiczny pozwala w sposób istotny ograniczać poziomy PEM w środowisku. W przypadku telewizji naziemnej zmiana z nadawania analogowego na cyfrowe pozwoliła nadawać z jednego nadajnika nawet do ośmiu programów dobrej jakości. W przypadku telefonii komórkowej po wydzieleniu ze stacji bazowej modułu radiowego RRU (Remote Radio Unit) i zintegrowaniu go z systemem antenowym z toru transmisyjnego usunięty został fider, który wnosił tłumienie około 4 -7 dB. Zmniejszenie tłumienia w torze od telefonu użytkownika do RRU stacji bazowej spowodowało, że poziom mocy emitowany z telefonu automatycznie zmniejszył się około 4 razy. Taka zmiana ma istotne znaczenie ponieważ telefon trzymany jest najczęściej blisko głowy.

2.3. Odległości zgodne z nowymi limitami PEM

Podstawowym dokumentem jaki należy przedstawić przy budowie lub modernizacji stacji bazowych jest kwalifikacja przedsięwzięcia. Obowiązek opracowania tego dokumentu wynika z § 2 ust. 1 pkt 7 i § 3 ust. 1 pkt 8 Rozporządzenia Rady Ministrów [9] z dnia 10 września 2019 r. Kwalifikacja polega na analizie położenia osi wiązki głównej z uwzględnieniem otoczenia stacji bazowej, takiego jak położenie i wysokość sąsiednich zabudowań, ukształtowanie terenu oraz długości wiązki głównej, którą porównuje się z długością podaną w rozporządzeniu [9] dla danej wartości mocy EIRP. Pomimo wprowadzonych zmian dopuszczalnych poziomów PEM w środowisku podane odległości w rozporządzeniu [9] nie zostały zmienione. Zmiany zasięgów stref zgodnych z nowymi limitami przedstawione są na rys. 2



Rys. 2. Długość wiązek głównych w osi głównej i odległości wypadkowej trzech składowych

Radiokomunikacyjne stacje bazowe emitują natężenia pól elektromagnetycznych o wartościach zmiennych w czasie i w przestrzeni oraz zależnych od natężenia ruchu jak również od zmiany w czasie lokalizacji terminali klientów. Rozporządzenie Ministra Klimatu [7] wskazuje metodę dotrzymywania poziomów PEM w środowisku polegającą na wykonywaniu pomiarów poprzez certyfikowane laboratoria.

3. LIMITY EKSPOZYCJI RF PEM

Systemy 5G wykorzystują częstotliwości, które są już uwzględnione w międzynarodowych standardach ochrony radiowej. Dwa organy międzynarodowe; Międzynarodowa Komisja ds. Ochrony przed Promieniowaniem Niejonizującym (ICNIRP) oraz Instytut Inżynierów Elektryków i Elektroników (IEEE) [4] opracowały wytyczne dotyczące narażenia i zdefiniowały limity narażenia w zakresie współczynnika absorpcji właściwej (SAR) oraz natężenia pola elektrycznego i magnetycznego oraz gęstości mocy. Te wytyczne dotyczące narażenia stanowią podstawę polityki i przepisów w wielu krajach.

Tab. 1. Międzynarodowe limity ekspozycji całego ciała na pole elektromagnetyczne (wg danych K. Suppl. 9-17)

Pasmo częstotliwości	*ICNIRP limit*	*Jednostka*	*IEEE limit C.95.1*
10 MHz – 3 GHz	0,08	W/kg	0,08
3 GHz – 10 GHz	0,08	W/kg	10 W/m^2
10 GHz– 300 GHz	10	W/m^2	10
400 MHz – 2 GHz	2 – 10	W/m^2	2 – 10
	(28 -61)	V/m	(27,5-61)
f > 2GHz	10	W/m^2	10
	(61)	V/m	(61)

W tabeli 1 przedstawiono podstawowe ograniczenia i poziomy odniesienia opracowane przez ICNIRP oraz IEEE C.95.1-2005. Średnie wartości graniczne mierzy się w ciągu sześciu minut. Pasma częstotliwości przydzielone do użytku przez 5G, w tym częstotliwości z zakresu 24 GHz-86 GHz, były od dziesięcioleci wykorzystywane do zestawiania łączy radioliniowych, do łączy satelitarnych i w technice radarowej. Poziomy PEM w środowisku 5G są mniejsze od limitów narażenia określonych przez Międzynarodową Komisję ds. Ochrony przed Promieniowaniem Niejonizującym. WHO, Komitet Naukowy ds. Pojawiających się i Nowo Rozpatrywanych Zagrożeń dla Zdrowia (SCENIHR) Unii Europejskiej oraz ICNIRP stwierdziły, że narażenie związane z sieciami bezprzewodowymi i ich użytkowaniem nie prowadzi do negatywnych skutków dla zdrowia, jeżeli jest poniżej limitów zalecanych przez ICNIRP. Badania nad możliwymi skutkami ekspozycji RF PEM na częstotliwościach powyżej 6 GHz na zdrowie człowieka sięgają wielu dziesięcioleci i są kontynuowane.

4. KLASY INSTALACJI STACJI BAZOWYCH

Uproszczone procedury oceny [ITU-T K100] służą do identyfikacji testowanego urządzenia EUT, o którym

wiadomo, że spełnia odpowiednie wartości graniczne narażenia, bez konieczności przestrzegania ogólnych lub kompleksowych procesów oceny narażenia. Jest to istotne na przykład ze względu na przesyłaną małą moc lub położenie nadajników lub anten i odpowiednich źródeł w stosunku do miejsc dostępnych dla ogółu społeczeństwa. Uproszczone procedury oceny opierają się na znajomości wartości poziomu mocy EIRP, charakterystyki anteny, instrukcji producenta oraz wysokości zainstalowania anteny w stosunku do miejsc dostępnych dla ludności w najbliższym otoczeniu.

4.1. Uproszczone procedury oceny zgodności urządzeń instalowanych w stacji bazowych

Uproszczone kryteria procedury oceny w celu spełnienia wymagań zgodności produktów instalowanych w stacji bazowej [3] i [2]:

E0 - Produkt jest zgodny z normą IEC 62479 lub wymiary brzegowe produktu są prawie zerowe. Brak szczególnych wymagań dotyczących instalacji produktu,

E2 – EIRP ≤ 2 W – Produkt jest instalowany zgodnie z instrukcjami producenta i/lub podmiotu oddającego do użytku. Zgodność z granicami narażenia są zwykle uzyskiwane przy zerowej odległości lub w odległości kilku centymetrów,

E10 – EIRP ≤ 10 W – Produkt jest instalowany zgodnie z instrukcjami producenta i/lub podmiotu wprowadzającego do eksploatacji i najniższa część promieniująca anteny lub anten znajduje się na wysokości co najmniej 2,2 m nad ogólnodostępnymi miejscami dla ogółu ludności,

E100 - EIRP ≤ 100 W- Produkt jest instalowany tak, że:

a) najniższa część promieniująca anteny znajduje się na wysokości co najmniej 2,5 m nad ogólnodostępnymi miejscami dla ogółu ludności.

b) minimalna odległość od obszarów dostępnych dla ogółu społeczeństwa w kierunku głównej wiązki wynosi D_m [m];

c) żadne inne źródła o częstotliwości radiowej o EIRP powyżej 10 W nie znajdują się w odległości 5 D_m [m] w kierunku głównego listka (określonej z uwzględnieniem szerokości wiązki połowy mocy) oraz w odległości D_m m w innych kierunkach. Jeżeli D_m nie jest dostępna, można zastosować wartość 2 m lub 1 m, jeżeli wszystkie częstotliwości nadawania produktu są równe lub wyższe niż 1 500 MHz.

E+ > 100 W – Produkt jest instalowany tak, że:

a) najniżej promieniująca część anteny (anten) znajduje się na minimalnej wysokości H_m [m] nad ogólnodostępnym miejscem dla ludności;

b) minimalna odległość od obszarów dostępnych dla ogółu społeczeństwa w kierunku wiązki głównej wynosi D_m [m];

c) żadne inne źródła o częstotliwościach radiowych, o EIRP powyżej 100 W nie znajdują się w odległości 5D_m [m] w wiązce głównej i w D_m [m] w innych kierunkach

Kryteria dotyczące wysokości montażu dla wartości EIRP ≤100 W zostały opracowane przy użyciu wzorów do oszacowania współczynnika absorpcji właściwej (SAR) podanego w [IEC 62232] w celu zapewnienia zgodności z podstawowymi ograniczeniami [1].

D_m to odległość zgodności w wiązce głównej. Jeśli odległość D_m nie jest dostępna, to można ją obliczyć za pomocą równań (2), (4), (6). Wysokość H_m określają równania (1), (3), (5).

Dla częstotliwości między 100 MHz a 400 MHz:

$$H_m = max \begin{cases} 2 + \sqrt{\frac{EIRP \cdot A_{sl}}{2 \cdot \pi}} \\ 2 + \sqrt{\frac{EIRP}{2 \cdot \pi}} \cdot \sin(\alpha + 1{,}129 \cdot \theta_{bw}) \end{cases} \quad (1)$$

$$D_m = \sqrt{\frac{EIRP}{2 \cdot \pi}} \quad (2)$$

Dla częstotliwości między 400 MHz a 2000 MHz:

$$H_m = max \begin{cases} 2 + \sqrt{\frac{EIRP \cdot 200 \cdot A_{sl}}{f \cdot \pi}} \\ 2 + \sqrt{\frac{200 \cdot EIRP}{f \cdot \pi}} \cdot \sin(\alpha + 1{,}129 \cdot \theta_{bw}) \end{cases} \quad (3)$$

$$D_m = \sqrt{\frac{EIRP \cdot 200}{f \cdot \pi}} \quad (4)$$

Dla częstotliwości między 2000 MHz a 40000 MHz

$$H_m = max \begin{cases} 2 + \sqrt{\frac{EIRP \cdot A_{sl}}{10 \cdot \pi}} \\ 2 + \sqrt{\frac{EIRP}{10 \cdot \pi}} \cdot \sin(\alpha + 1{,}129 \cdot \theta_{bw}) \end{cases} \quad (5)$$

$$D_m = \sqrt{\frac{EIRP}{10 \cdot \pi}} \quad (6)$$

gdzie:
f- częstotliwość z dolnego zakresu danego pasma [MHz]
A_{sl} – wartość tłumienia listków bocznych
α – kąt pochylenia wiązki głównej (tilt) w radianach
θ_{bw} – pionowa szerokość wiązki o połowie mocy w radianach.

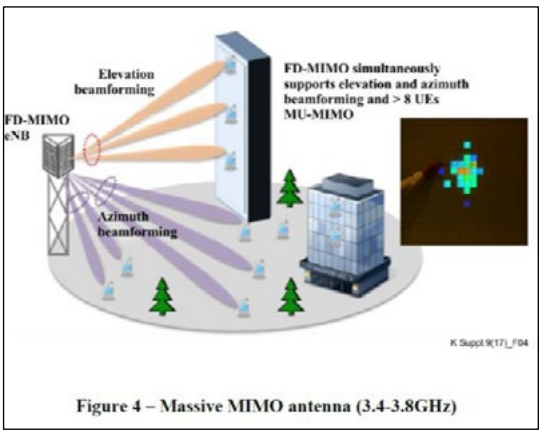
4.2. Pasma częstotliwości dla 5G

Systemy 5G będą wykorzystywać znacznie gęściejsze sieci z dużą liczbą mikrostacji bazowych, zlokalizowanych znacznie bliżej użytkowników [4]. Wewnątrz budynku może znajdować się wiele wewnętrznych mikrostacji bazowych. Na mniej zaludnionych obszarach nadal będą wdrażane zarówno makrostacje bazowe, jak i małe komórki, więc sieć będzie wysoce heterogeniczna. 5G potrzebuje widma w trzech kluczowych zakresach częstotliwości, aby zapewnić szeroki zasięg i obsługiwać wszystkie planowane usługi:
1. Niskie pasmo (poniżej 1 GHz) — zapewniające szeroki zasięg na obszarach miejskich, podmiejskich i wiejskich oraz obsługujące IoT dla aplikacji o niskiej szybkości transmisji danych,

2. Średnie pasmo (1-6 GHz) – zapewniające dobry zasięg i wysokie prędkości, w tym oczekiwany początkowy zakres 5G 3,4-3,8 GHz, który został zidentyfikowany jako najbardziej prawdopodobne pasmo do uruchomienia 5G na całym świecie,
3. Wysokie pasmo (powyżej 6 GHz) – zapewnia ultrawysokie prędkości szerokopasmowe dla zaawansowanych mobilnych aplikacji szerokopasmowych i które jest najbardziej odpowiednie do zastosowań w miejscach o dużym natężeniu ruchu. Pasmo 26-28 GHz zostało zidentyfikowane przez niektóre administracje dla przyszłych zastosowań 5G.

4.3. Określenie rzeczywistej mocy maksymalnej dla ocen zgodności RF-PEM sieci 5G strony

Postępy w technologii sieci bezprzewodowych dla 5G spowodowały, że sieci bezprzewodowe stały się znacznie bardziej wydajne i wymagają mniejszej mocy transmisji, aby zapewnić te same szybkości transmisji danych (ITU-T K Supplement 16) [5]. Sieci 5G będą transmitować podobne poziomy mocy w porównaniu z poprzednimi technologiami mobilnymi. Podobnie jak obecne sieci 2G, 3G i 4G, stacje bazowe 5G nie będą projektowane do pracy z maksymalną mocą, z wyjątkiem bardzo krótkich okresów w celu obsługi wahań ruchu. Oznacza to, że transmitowana moc jest uśredniana w okresach czasu istotnych dla oceny narażenia na RF-PEM, np. sześć minut, jest znacznie niższa niż maksymalna moc znamionowa stacji bazowej. W konsekwencji użycie maksymalnej mocy znamionowej doprowadzi do zbyt konserwatywnych (przeszacowanych) wartości ekspozycji RF-PEM i granic zgodności, zwłaszcza w przypadku współistniejących kilku różnych technologii i anten. Aby rozwiązać ten problem, zarówno rekomendacja ITU-T K.100 [3], jak i norma IEC 62232 [2] otwierają możliwość wykorzystania „rzeczywistej mocy maksymalnej", którą można określić na podstawie pomiarów rzeczywistej mocy wyjściowej stacji bazowej, na podstawie pomiarów dużej liczby reprezentatywnych stacji bazowych w sieci lub za pomocą modeli statystycznych lub symulacji sieciowych. Rzeczywistą moc maksymalną można na przykład przyjąć jako wartość 95 percentyla uzyskanego rozkładu mocy [b-IEC TR 62669].



Rys. 3. Stacja 5G z anteną Massive MIMO (źródło K. Suppl.9 2017)

W przypadku oceny narażenia na pola elektromagnetyczne w lokalizacjach 5G przy użyciu mMIMO ważne jest dokładne określenie rzeczywistej maksymalnej transmitowanej mocy. Stacje bazowe Massive MIMO, takie jak przedstawiona na Rys. 3, przesyłają jednocześnie wiele wiązek do podłączonych urządzeń. Wiązki te zmieniają się szybko zarówno w czasie, jak i w przestrzeni, i przez długi czas nie będzie transmisji w określonym kierunku przy maksymalnej mocy znamionowej. Raport techniczny [b-IEC TR 62669] zawiera szczegółowe wskazówki dotyczące określania rzeczywistej mocy maksymalnej dla anten mMIMO. (Raport techniczny b-IEC TR 62669 zawiera serię studiów przypadków, w których ocenia się pola elektromagnetyczne zgodnie z normą IEC 62232:2017).

4.4 Wyniki ocen 5G RF-PEM

Na całym świecie krajowe agencje ds. zdrowia, rządowe organy regulacyjne, środowiska akademickie, laboratoria badawcze, operatorzy komórkowi i producenci przeprowadzili szeroko zakrojone testy w sieciach komercyjnych i testowych w celu określenia poziomów narażenia na promieniowanie elektromagnetyczne 5G [5].

Interaktywna mapa zawiera dane pochodzące od wielu interesariuszy. Zmierzone poziomy dotyczą obszarów publicznie dostępnych, zwykle na poziomie gruntu. Poziomy RF-PEM 5G zostały przeliczone na procent odpowiedniego międzynarodowego [b-ICNIRP] publicznego limitu RF-PEM.
Z podsumowania wynika, że:

1. obowiązują międzynarodowe standardy bezpieczeństwa i testowania dla wszystkich częstotliwości 5G, w tym fal milimetrowych;

2. zmierzone poziomy z sieci 5G działających na wszystkich kontynentach są niskie i znacznie poniżej międzynarodowych limitów bezpieczeństwa;

3. Poziomy RF-PEM 5G są podobne do innych technologii bezprzewodowych, z niewielką różnicą między pasmami częstotliwości;

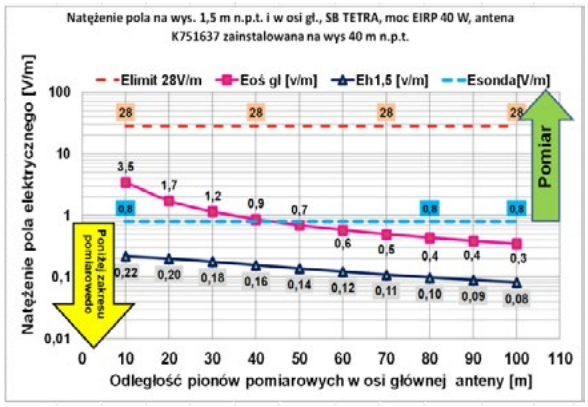
4. Typowy maksymalny zmierzony poziom 5G RF-PEM w badaniach wynosi około lub mniej niż 1% (gęstość mocy) międzynarodowych ograniczeń publicznych.

5. Sieci 5G są zaprojektowane tak, aby efektywnie wykorzystywać widmo radiowe, co widać w testach. Zmniejszone wymagania sygnalizacyjne dla 5G i wykorzystanie inteligentnych systemów antenowych oznaczają bardziej efektywne wykorzystanie energii RF-PEMF.

5. PODSUMOWANIE

Norma PN-EN 62232:2018 otwiera możliwości wyznaczania gęstości mocy i natężenia pola wokół instalacji stacji bazowych pracujących w zakresie od 100 MHz do 300 GHz w celu zachowania zgodności z dopuszczalnymi limitami. W zakresie częstotliwości 148 MHz -175 MHz pracują w Polsce stacje bazowe z mocą EIRP około 25 W i w zakresie 380 MHz - 450 MHz z mocą EIRP około 40W. Stacje bazowe pracujące w tych zakresach i z taką mocą kwalifikują się do uproszczonej procedury oceny zgodności instalacji stacji bazowych. Obecnie nie można skorzystać z tego udogodnienia, ponieważ na podstawie załącznika do Rozporządzenia [7] ust.1 pkt 1 ustawo-

dawca wymaga, aby prowadzący instalację zlecił wykonanie pomiarów PEM. Podstawę stanowi dogmat niepodlegający sprawdzeniu zawarty w art. 122a POŚ[10] o instalacjach mogących negatywnie oddziaływać z powodu mocy EIRP ≥ 15 W. Na Rys. 4. jest przedstawiony rozkład natężenia pola wokół SB TETRA. W miejscach dostępnych dla ludności wartość natężenia pola wynosi 0,22 V/m, a dolny zakres pomiarowy wynosi 0,8 V/m. Wskaźnik $(E_h^2 / E^2_{limit}) = 0{,}00006$.



Rys. 4. Rozkład natężenia pola wokół stacji bazowej TETRA pracującej z mocą EIRP 40 W.

W dalszym postępowaniu ustawodawca wymaga, aby inwestor na podstawie tego samego zapisu o mocy EIRP ≥ 15 W dokonał zgłoszenia instalacji. Aby korzystać w pełni z metod zawartych w normie PN-EN 62232 do wyznaczania natężenia pola, gęstości mocy w otoczeniu stacji bazowych każdego typu, w tym również stacji z anteną Massime MIMO konieczne są istotne zmiany w Ustawie Prawo Ochrony Środowiska[10], w Rozporządzeniu Ministra Klimatu [7] oraz w innych rozporządzeniach, które nie zostały uaktualnione pomimo wprowadzenia nowych limitów PEM Rozporządzeniem Ministra Zdrowia[8].

LITERATURA

[1] ICNIRP *Guidelines for limiting exposure to electromagnetic fields (100 kHz-300 GHz),* Health Phys, March 2020

[2] IEC 62232:2017, Determination of RF field trength, power density and SAR in the vicinity of radiocommunication base stations for the purpose of evaluating human exposure

[3] Rec. ITU K.100 2021, Measurement of radio frequency electromagnetic fields to determine compliance with human exposure limits when a base station is put into service

[4] ITU-T K Supplement 9 2017, 5G technology and human exposure to RF EMF

[5] ITU-T K Supplement 16 2022, Electromagnetic field compliance assessments for 5G wireless network

[6] PN-EN 62232:2018, Wyznaczanie natężenia pola RF, gęstości mocy i SAR w otoczeniu radiokomunikacyjnych stacji bazowych dla oceny poziomu ekspozycji człowieka

[7] Rozporządzenie Ministra Klimatu z dnia 17 lutego 2019 r., w *sprawie sposobów sprawdzania dotrzymywania dopuszczalnych poziomów pól elektromagnetycznych w środowisku,* Dz. U. Nr 2020 poz.258

[8] Rozporządzenie Ministra Zdrowia z dnia 17 grudnia 2019 r., *w sprawie dopuszczalnych poziomów pól elektromagnetycznych w środowisku,* Dz. U. 2019 poz. 2448

[9] Rozporządzenie Rady Ministrów z dnia 10 września 2019 r., *w sprawie przedsięwzięć mogących znacząco oddziaływać na środowisko,* Dz. U. 2019 poz. 1839

[10] Ustawa z dnia 27 kwietnia 2001 r., *Prawo ochrony środowiska,* Dz. U. 2022 poz. 2556

ADAPTACYJNE KSZTAŁTOWANIE CHARAKTERYSTYKI PROMIENIOWANIA W MILIMETROWYCH SIECIACH 5G

ADAPTIVE BEAMFORMING IN MM-WAVE 5G NETWORKS

Jerzy Martyna
Wydział Matematyki i Informatyki UJ
ul. Prof. S. Łojasiewicza 6, 30-348 Kraków
e-mail: jerzy.martyna@uj.edu.pl

DOI: 10.15199/59.2023.4.87

Streszczenie: **W artykule przedstawiono zastosowanie podejścia opartego o uczenie maszynowe dla rozwiązania problemu kształtowania charakterystyki promieniowania (ang. *beamforming*, BF) w milimetrowych sieciach 5G. W procesie modelowania BF wykorzystano problem wielorękich bandytów, w tym algorytm $Exp3$. Dzięki jego użyciu znaleziono przybliżone rozwiązanie tego problemu. Wyniki symulacji potwierdziły, że proponowany schemat jest porównywalny pod względem efektywności energetycznej ze znanymi rozwiązaniami.**

Abstract: **The article presents the use of a machine learning approach for the solution *beamforming*, BF) in millimeter 5G networks. The multi-armed bandit problem, including the $Exp3$ algorithm, was used in the BF modeling process. Thanks to its use found an approximate solution to this problem. The simulation results confirmed that the proposed scheme is comparable in terms of energy efficiency with known solutions.**

Słowa kluczowe: **Milimetrowe sieci komórkowe 5G, kształtowanie charakterystyki promieniowania, uczenie maszynowe, problem wielorękich bandytów.**

Keywords: **mm-wave 5G network, beamforming, machine learning, beam management, multi-armed bandit problem.**

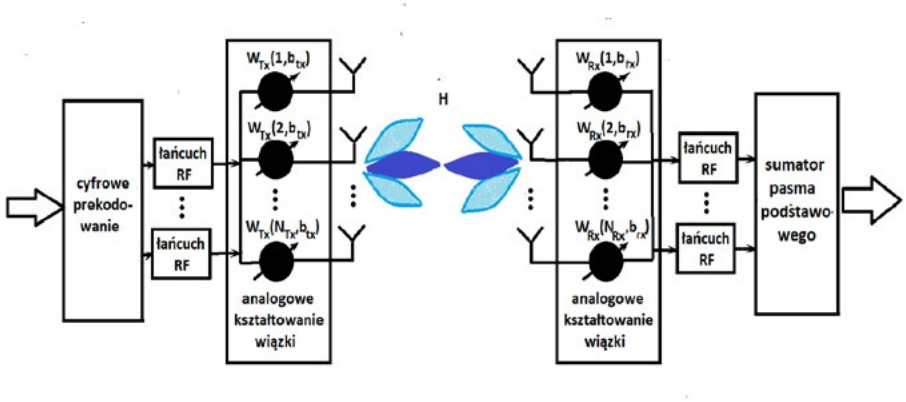
1. WPROWADZENIE

Stale wzrastające zapotrzebowanie na duże ilości danych cyfrowych, jak również niedobór częstotliwości używanych w urządzeniach radiokomunikacji ruchomej, są wyzwaniem dla projektantów systemów komórkowych 5G. Sposobem na przezwyciężenie tych problemów jest wykorzystanie fal milimetrowych opartych o nieużywane dotąd częstotliwości (30 - 300 GHz). Wadą tego rodzaju systemów jest redukcja rozmiaru komórek [1], jak również potrzeba stosowania wąskich, ukierunkowanych wiązek (ang. *beams*) [2], co wiąże się ze śledzeniem wiązek, zarządzaniem mobilnością, nowymi procedurami przenoszenia połączeń itp.

Podstawą sieci komórkowych 5G i ponad 5G (B5G) są nowe technologie i koncepcje, w tym technologia inteligentnych anten MIMO (ang. *massive MIMO*, mMIMO), kształtowanie charakterystyki promieniowania (ang. *beamforming*, BF), schemat nieortogonalnej metody wielodostępu (NOMA) itp. System mMIMO [3] umożliwia jednoczesną obsługę wielu użytkowników z wykorzystaniem tych samych zasobów czasowo-częstotliwościowych, podczas, gdy w konwencjonalnym systemie antenowym wielu użytkowników korzysta z jednego zasobu widmowego o ustalonej szerokości. Umożliwia to kilkukrotne zwiększenie pojemności sieci. Użycie techniki kształtowania charakterystyki promieniowania jest sposobem dostarczania sygnału radiowego tylko w kierunku określonego użytkownika korzystającego z zasobów radiowych i w takiej ilości jaka jest wymagana. Z kolei metoda NOMA oferuje najwyższą pojemność sumaryczną, oferując przy tym użytkownikom możliwość transmisji dużych ilości danych [4].

Wiele publikacji poświęcono zagadnieniom kształtowania charakterystyki promieniowania. Między innymi, w pracy Gotsisa i in. [5] podzielono kształtowanie wiązki promieniowania na dwie główne kategorie: przełączane kształtowanie wiązki i adaptacyjne kształtowanie wiązki. Ponadto, podzielono te techniki na różne typy w zależności od użycia anten macierzowych, liniowych, kołowych, i prostokątnych. W artykułach [6] oraz [7] sklasyfikowano technologię BF na następujące rodzaje: analogowe i cyfrowe kształtowanie wiązki oraz hybrydowe, czyli analogowo-cyfrowe kształtowanie wiązki.

Hybrydowe kształtowanie wiązki [8] pozwala na obsługę wielu strumieni transmisji z multipleksowaniem przestrzennym, jak również przestrzennym podziałem. Osiąga ono wydajność spektralną porównywalną do w pełni cyfrowego kształtowania wiązki, przy jednoczesnym zredukowaniu złożoności sprzętowej. W związku z tym technika ta jest uznawana za bardzo obiecującą metodę kształtowania wiązek w sieciach milimetrowych 5G. Użycie metod sztucznej inteligencji do kształtowania wiązek w tej metodzie doczekało się wielu przykładów. Między innymi, w pracy [9] zaproponowano schemat adaptacji łącza wspomagany uczeniem maszynowym dla sieci z hybrydowym kształtowaniem wiązki.



Rys. 1. Hybrydowe kształtowanie par wiązek w milimetrowej sieci 5G.

Celem pracy jest sformułowanie zagadnienia kształtowania charakterystyki promieniowania w milimetrowych sieciach 5G z hybrydowym kształtowaniem wiązki jako problemu wielorękich bandytów. Następnie przy użyciu algorytmu $Exp3$ znaleziono przybliżone rozwiązanie zadania wyboru par wiązek maksymalizującego osiągalną przepustowość transmisji. Opracowana metoda uwzględnia się adaptacyjne kształtowanie wiązki uwzględniająca ruch urządzenia.

Treść pracy jest następująca. W rozdziale 2 przedstawiono przyjęty model systemu. Rozdział 3 opisuje metodę adaptacyjnego kształtowania charakterystyki promieniowania w milimetrowej sieci 5G opartą o algorytm wielorękiego bandyty znany jako $Exp3$. W rozdziale 4 opisano szczegóły implementacyjne proponowanej metody. Rozdział 5 zawiera wyniki badań symulacyjnych. W ostatnim rozdziale podsumowano pracę.

2. MODEL SYSTEMU

Na rys. 1 przedstawiono model systemu dla hybrydowego kształtowania wiązki złożony z dwóch anten: nadawczej N_{Tx} oraz odbiorczej N_{Rx}. Przyjęto, że $W(n,b)$ jest wektorem wagi anteny (ang. *antenna weight vector*, AWV) n-tego elementu anteny wymaganego dla sterowania kierunku b. Macierze $\mathbf{W}_{Tx}$ oraz $\mathbf{W}_{Rx}$ reprezentują tutaj liczbę, rozmiar oraz rozstaw elementów anteny. Dla anten fazowanych kolumny macierzy książki kodowej (ang. *codebook matrix*) określają przesunięcie fazowe używane dla poszczególnych elementów anteny dla kształtowania wiązki w określonym kierunku.

W oparciu o macierze $\mathbf{W}_{Tx}$ oraz $\mathbf{W}_{Rx}$ dla danego sygnału $\mathbf{x}$ odebrany sygnał $\mathbf{y}$ może być wyrażony w następujący sposób [12]:

$$\mathbf{y} = \mathbf{W}_{Rx}^{H}(:,b_{rx})\mathbf{H}\mathbf{W}_{Tx}(:,b_{tx})\mathbf{x} + \mathbf{W}_{Rx}(:,b_{rx})\mathbf{n} \quad (1)$$

gdzie $\mathbf{W}_{Tx}(:,b_{tx})$ oraz $\mathbf{W}_{Rx}(:,b_{rx})$ są wektorami AWV odpowiednio dla Tx oraz Rx o długościach $N_{Tx} \times 1$ oraz $N_{Rx} \times 1$ w kierunkach b_{tx} oraz b_{rx}. H określa tutaj transpozycję hermitonowską oraz $\mathbf{n}$ wektorem addytywnego białego szumu gaussowskiego (ang. *additive white gaussian noise*, AWGN) o długości $N_{Rx} \times 1$.

Macierz kanałowa $\mathbf{H}$ typu $N_{Rx} \times N_{Tx}$ może być przedstawiona w postaci:

$$\mathbf{H} = \frac{1}{\delta}\sum_{l=1}^{L}\gamma_l \mathbf{T}_{Rx}(\Phi_l)\mathbf{T}_{Tx}^{H}(\Theta_l) \quad (2)$$

gdzie l, $1 \le l \le L$ jest indeksem ścieżki kanałowej, L jest całkowitą liczbą ścieżek, δ jest współczynnikiem średniej utraty ścieżki. (zależnym od rozstawu pomiędzy Tx oraz Rx, γ_l jest zyskiem ścieżki w ścieżce kanałowej l, $\Phi_l \in [0, 2\pi]$ oraz $\Theta_l \in [0, 2\pi]$ są odpowiednio kątami AoA (Angle of Arrival) oraz AoD (Angle of Departure) dla ścieżki l.

Wektory $\mathbf{T}_{Tx}(\Theta_l)$ oraz $\mathbf{T}_{Rx}(\Phi_l)$ mogą być odpowiednio przedstawione w następujący sposób:

$$\mathbf{T}_{Tx}(\Theta_l) = [1, e^{j\frac{2\pi}{\lambda}d\sin(\Theta_l)}, \ldots, e^{j(N_{Tx}-1)\frac{2\pi}{\lambda}d\sin(\Theta_l)}]^T \quad (3)$$

$$\mathbf{T}_{Rx}(\Phi_l) = [1, e^{j\frac{2\pi}{\lambda}d\sin(\Phi_l)}, \ldots, e^{j(N_{Rx}-1)\frac{2\pi}{\lambda}d\sin(\Phi_l)}]^T \quad (4)$$

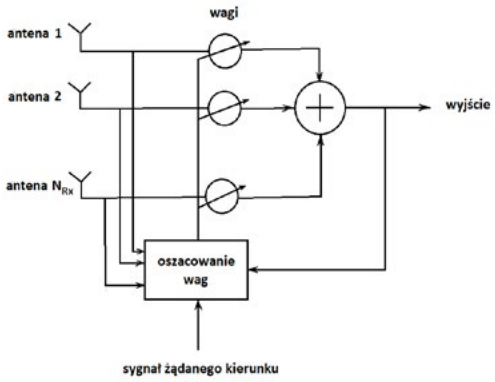
gdzie λ jest długością fali milimetrowej, d jest dystansem pomiędzy Tx a Rx.

2.1. Zagadnienie optymalizacji formowania wiązki w sieci milimetrowej 5G

Problem optymalizacji formowanej wiązki w sieci milimetrowej 5G polega na znalezieniu najlepszej pary wiązek (b_{tx}^*, b_{rx}^*), która maksymalizuje osiągalną przepustowość transmisji danych w określonym kierunku [13]. Można to przedstawić w postaci:

$$(b_{tx}^*, b_{rx}^*) = W\left(\frac{T_D}{ZT_{BT}+T_D}\right)$$
$$\arg\max_{\forall(b_{tx},b_{rx})} log_2\left(1 + \frac{|\mathbf{F}_{Rx}^{H}(:,b_{rx})\mathbf{H}\mathbf{F}_{Tx}(:,b_{tx})|^2}{\alpha_0}\right) \quad (5)$$

przy czym $b_{tx} \in \Phi_{BTx}$, $b_{rx} \in \Phi_{BRx}$, $F_{Tx} \in \Phi_{CTx}$, $F_{Rx} \in \Phi_{CRx}$.



Rys. 2. Adaptacyjne kształtowanie wiązki

Użyta w równaniu (5) wielkość W określa szerokość osiągalnego pasma, α_0 jest mocą szumu, Φ_{BTx}, Φ_{BRx}, Φ_{CTx}, Φ_{CRx} są przestrzeniami kierunku wiązki odpowiednio dla Tx, Rx, $AWV\ Tx$ oraz $AWV\ Rx$. Iloraz $\frac{T_D}{(ZT_{BT}+T_D)}$ przedstawia narzut czasowy związany z uczeniem się algorytmu, przy czym T_D jest czasem transmisji danych, T_{BT} jest czasem uczenia się kształtowania wiązki, Z jest całkowitą liczbą par wiązek (b_{Tx}, b_{Rx}) użytych w procesie uczenia.

2.2. Adaptacyjne kształtowanie wiązki

Adaptacyjne kształtowanie wiązki pozwala na dostosowanie jej do zmieniającego kierunku odbieranego sygnału, wynikającego m.in. z ruchu użytkownika. Jedną z metod stosowanych do adaptacyjnego kształtowania wiązki jest technika oparta o minimalizację błędu średniokwadratowego (ang. *Least Mean Squares*, LMS), opisaną po raz pierwszy przez Widrowa i in. [14]. Jej działanie polega na iteracyjnej aktualizacji wag związanych z antenami (patrz rys. 2) poprzez oszacowanie gradientu błędu średniokwadratowego, a następnie przesunięciu tych wag w ujemnym kierunku gradientu proporcjonalnie do przyjętego parametru μ. Układ jest stabilny, jeśli jest spełniony warunek [15]:

$$0 \leq \mu \leq \frac{1}{2\Lambda_{max}} \tag{6}$$

gdzie Λ_{max} jest największą wartością własną macierzy korelacji $\mathbf{R}_{xx}(k)$ w k-tym kroku podanej w postaci:

$$\mathbf{R}_{yy}(k) = \mathbf{y}(k)\mathbf{y}^H(k) \tag{7}$$

gdzie $\mathbf{y}(k)$ określa wartość odbieranego sygnału. Wektor sygnału błędu jest podany następująco:

$$\mathbf{e}(k) = \mathbf{D}(k) - \mathbf{w}^H(k)\mathbf{y}(k) \tag{8}$$

gdzie $\mathbf{D}(k)$ jest wektorem pożądanych wartości sygnału.

Algorithm 1 Adaptacyjne kształtowanie charakterystyki promieniowania w milimetrowych sieciach 5G

procedure *Formowanie par wiązek*
Input: W_{Tx}, W_{Rx}, γ
Output: (b^*_{tx}, b^*_{rx})
Initiation: for $m \leftarrow 1, M$ **do** $R_{m,t} = 0$, $w_m(1) = 1$;
  for $t \leftarrow 1, T$ **do**
    for $m \leftarrow 1, M$ **do**
      ustaw $p_m(t) = (1-\gamma)\frac{w_m(t)}{\sum_{j-1}^{M} w_j(t)} + \frac{\gamma}{M}$;
      weźć następną wartość m_t wybraną losowo zgodnie z rozkładem prawd. $p_m(t)$;
      weźć parę wiązek i oblicz nagrodę $R_{m,t}$
      $(b^*_{tx}, b^*_{rx})_{m,t} = \arg\max \sum_{m=1}^{M} E[R_{m,t}]$;
      oblicz szacowaną nagrodę $\hat{R}_{m,t} = \frac{R_{m,t}}{p_m(t)}$;
      ustaw wagę $w_m(t+1) = w_m(t)e^{\frac{\gamma \hat{R}_{m,t}}{M}}$;
      ustaw pozostałe wagi $w_j(t+1) = w_j(t)$;
    end for
  end for
end procedure;

for $k \leftarrow 1, K_{max}$ **do**
  repeat
    Formowanie par wiązek;
    oblicz macierz kanałową $\mathbf{H}$;
    oblicz macierz korelacji $\mathbf{R}_{yy}(k)$;
    oblicz wartości własne Λ_{max};
    oblicz $\mathbf{e}(k)$;
    oblicz $\mathbf{w}(k+1)$;
  until $(\mu < 0)$ **or** $(\mu > \frac{1}{2\Lambda_{max}})$
end for

Rys. 3. Pseudokod programu adaptacyjnego kształtowania wiązki.

Wektor wag $\mathbf{w}$ jest aktualizowany w kroku $(k+1)$ zgodnie z równaniem:

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \mu\mathbf{e}^*(k)\mathbf{y}(k) \tag{9}$$

Znak $*$ oznacza tu sprzężoną transpozycję.

3. IMPLEMENTACJA ADAPTACYJNEGO FORMOWANIA WIĄZKI W MILIMETROWYCH SIECIACH 5G

W tej części przedstawia się implementację nowej techniki adaptacyjnego formowania wiązki w milimetrowych sieciach 5G. Pseudokod tej metody przedstawiono na rys. 3. Procedura *Formowanie par wiązek* umożliwia rozwiązanie problemu optymalizacji sformułowanego w rów. (5). Wykorzystano tutaj jeden z algorytmów wielorękich bandytów [16], który jest znany jako $Exp3$ [17]. Algorytm ten wybiera pojedynczą parę wiązek, maksymalizując osiągalną przepustowość transmisji.

Tablica 1. Główne parametry symulacji.

Parametr	*Wartość*
moc stacji bazowej	30 [dBm]
promień komórki	300 [m]
szerokość pasma systemu	10 [GHz]
liczba UE	1, 2, 3
szerokość wiązki	$10^o, 20^o, 30^o, 40^o, 50^o$
T_D	1 [ms]
T_{BT}	25 [ms]
ϵ	0.1
T	1000 [s]

Działanie algorytmu $Exp3$ w odniesieniu do formowania wiązek w sieciach milimetrowych 5G rozpoczyna się od inicjacji, polegającej na wyborze pary wiązek dla M sygnałów radiowych oraz obliczania odpowiadających im nagród (ang. *reward*) dla tych sygnałów. które można przedstawić w postaci:

$$R_{m,l} = \frac{T_D}{ZT_{BT} + T_D} \log_2\left(1 + \frac{|F_{Rx}^H(:, b_{Tx,m,t}) H F_{Tx}(:, b_{tx,m,t})|^2}{\alpha_0}\right), \quad 1 \le m \le M, 1, \le t \le T \tag{10}$$

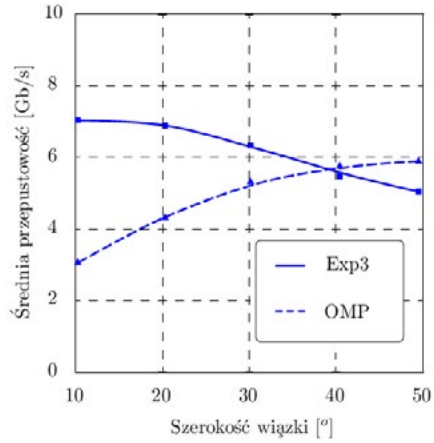
gdzie $(b_{tx,m,t}, b_{rx,m,t})$ jest wybraną parą wiązek m, $1 \le m \le M$ w chwili t, T jest przyjętym horyzontem czasu. Złożoność obliczeniowa algorytmu $Exp3$ jest równa co najwyżej $O(\sqrt{MT\log(M)})$.

Po inicjacji algorytm $Exp3$ utrzymuje listę wag dla każdej akcji, a następnie przy użyciu tych wag dokonuje on wyboru następnej akcji. Przyjęty tu czynnik egalitarny γ, $\gamma \in [0, 1]$ pozwala na dostosowanie losowo wybranej akcji do stawianych wymagań. Jeśli $\gamma = 1$, to wtedy wagi nie wpływają na decyzję wyboru w żadnym kroku algorytmu.

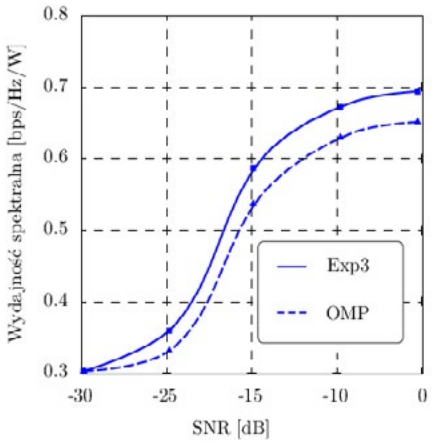
W części głównej programu (por. rys. 3) nastepuje iteracyjne wywołanie procedury *Formowanie par wiązek* dla znalezienia pary wiązek w k-tym kroku. Następnie kolejno są obliczane dla każdego kroku k: macierz kanałowa $\mathbf{H}$ (rów. 2), macierz korelacji $\mathbf{R}_{yy}$ (rów. 7), Λ_{max} - największa wartość własna macierzy korelacji dla macierzy korelacji $\mathbf{R}_{yy}$, wektor sygnału błędu $\mathbf{e}$ (rów. 8) oraz wektor wag przypisywanych antenom dla następnego kroku (rów. 10). Warunkiem działania opisanej tu pętli głównej programu jest spełnienie warunku stabilności układu podanego przez rów. 6.

4. WYNIKI BADAń SYMULACYJNYCH

W tej części zostaną przedstawione wyniki badań symulacyjnych dla badania efektywności adaptacyjnego kształtowania wiązki w milimetrowych sieciach 5G. Przyjęto, że system stanowi pojedyncza komórka z pojedynczą stacją bazową oraz zmieniającą się liczbą urządzeń zewnętrznych (ang. *user equipment*, UE). Dla uproszczenia obliczeń założono identyczną liczbę ścieżek kanałowych pomiędzy UE a stacją bazową. Podstawowe parametry symulacji przedstawiono na w tabeli 1.



Rys. 4. Średnia przepustowość systemu w zależności od szerokości wiązki.



Rys. 5. Wydajność spektralna w zależności od SNR.

Na rys. 4 przedstawiono średnie przepustowości proponowanej metody w zależności od szerokości kąta wiązki. Wyniki te porównano z rezultatami uzyskanymi przy użyciu algorytmu hybrydowego kształtowania wiązki określanego mianem OMP (Orthogonal Matching Pursuit) [18]. Z przedstawionych wykresów wynika, że proponowana metoda jest bardziej efektywna dla mniejszych wartości szerokości wiązki.

Na rys. 5 pokazano wydajność spektralną (ang. *spectral efficiency*) proponowanej metody w zależności od współczynnika SNR dla obydwóch metod kształtowania wiązek. Jest widoczne, że zastosowanie metody uczenia maszynowego w doborze wiązek w sieciach milimetrowych 5G jest efektywniejsze od konwencjonalnego podejścia.

5. WNIOSKI

W artykule wprowadzono metodę adaptacyjnego kształtowania wiązki w milimetrowych sieciach 5G. O wyborze nowej pary wiązek decyduje w niej algorytm $Exp3$, który

użyto do przybliżonego rozwiązania problemu wielorękich bandytów. Samą adaptację wiązki uzyskuje się dzięki minimalizacji błędu średniokwadratowego. Dla przyjętego scenariusza przeprowadzono symulację skuteczności opracowanej metody, jak i określeniu jej efektywności energetycznej.

Literatura

[1] Rangan S., Rappaport T.S., Erkip E. 2014. Millimeter-wave cellular wireless networks: potentials and challenges, Proc. of the IEEE, 102(3) : 366 - 385.

[2] Roh W., Seol J.Y., Park J., Lee B., Lee J., Kim Y., Cho J., Cheun K., Aryanfar F. 2014. Millimeter-wave beamforming as an enabling technology for 5G cellular communications: theoretical feasibility and prototype results, IEEE Communication Magazine, 52(2) : 106 - 113.

[3] Samsung, Massive MIMO for New Radio, 2020. [Online]. Available: https://www.samsung.com/global/business/networks/insights/white-papers/1208-massive-mimo-for-new-radio/

[4] Liu Y., Qin Z., Elkashlan M., Ding Z., Nallanathan A., Hanzo L. 2017. Nonorthogonal Multiple Access for 5G and Beyond, Proceedings of the IEEE, 105(12): 2347 - 2381.

[5] Gotsis K.A., Sahalos J.N. 2011. Beamforming in 3G and 4G mobile communications: the switched-beam approach. In: Maícas, J.P. (Ed.), A Multidisciplinary Approach. InTech, 201 - 216.

[6] Hur, S., Kim, T., Love, D.J., Krogmeier D.J., Thomas J.V., Ghosh T.A. 2013. Millimeter wave beamforming for wireless backhaul and access in small cell networks. IEEE Transactions on Communications, 61(10) : 4391 - 4403.

[7] Bogale T.E., Le L.B., Haghighat A., Vandendorpe L. 2016. On the number of RF chains and phase shifters, and scheduling design with hybrid analog–digital beamforming. IEEE Transactions on Wireless Communications, 15(5) : 3311 - 3326.

[8] Zhang J., Yu X., Letaief K.B. 2020. Hybrid Beamforming for 5G and Beyond Millimeter-Wave Systems: A Holistic View, IEEE Open Journal of the Communications Society, 1 : 77 - 91.

[9] Satyanarayana K., El-Hajjar M., Mourad A.A., Hanzo L. 2019. Multi-user hybrid beamforming relying on learning-aided link-adaptation for mmWave systems. IEEE Access, 7 : 23197 - 23209.

[10] Chen, J., Feng, W., Xing, J., Yang, P., Sobelman, G. E., Lin, D., Li, S. (2020). Hybrid beamforming/combining for millimeter wave MIMO: A machine learning approach. IEEE Transactions on Vehicular Technology, 69(10) : 11353 - 11368.

[11] Hojatian H., Nadal J., Frigon J. F., Leduc-Primeau F. 2021. Unsupervised deep learning for massive MIMO hybrid beamforming. IEEE Transactions on Wireless Communications, 20(11) : 7086 - 7099.

[12] Long Y., Chen Z., Fang J., Tellambura C. 2018. Data-driven-based analog beam selection for hybrid beamforming under mm-wave channels. IEEE Journal of Selected Topics in Signal Processing, 12(2) : 340 - 352.

[13] Gatzianas, M., Kalfas, G., Vagionas, C., Mesodiakaki, A. 2019. Downlink coordinated beamforming policies for 5G millimeter wave dense networks. In 2019 European Conference on Networks and Communications (EuCNC) : 342 - 346.

[14] Widrow B., Manty P.E., Griffiths L.J., Goode B.B. 1967. Adaptive Antenna Systems, Proc. IEEE, 55(12) : 2143 - 2159.

[15] Monzingo R., Miller T. 1980. Introduction to Adaptive Arrays, Wiley Interscience,, John Wiley and Sons, New York.

[16] Auer P., Cesa-Bianchi N., Fischer P. 2002. Finite-time Analysis of the Multiarmed Bandit Problem, Machine Learning, (47) : 235 - 256.

[17] Seldin Y., Szepesvári C., Auer P., Abbasi-Yadkori Y. 2012. Evaluation and Analysis of the Performance of the EXP3 Algorithm in Stochastic Environments. In 10th European Workshop on Reinforcement Learning, 103 - 116.

[18] Bogale T.E., Le L.B. 2014. Beamforming for multiuser massive MIMO systems: Digital versus hybrid analog-digital. In 2014 IEEE Global Communications Conference (GLOBECOM) : 4066-4071.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

ADAPTACJA DYNAMICZNEGO ROZDZIAŁU ZASOBÓW W SYSTEMIE KOMUNIKACJI V2X W TRYBIE 1 W SIECIACH 5G NEW RADIO

RESOURCE ADAPTATION IN 5G NR V2X MODE 1 DYNAMIC RESOURCE ALLOCATION

Saif Sabeeh[1]; Krzysztof Wesołowski[2];

[1] Politechnika Poznańska, Poznań, saifsadeq1982@gmail.com
[2] Politechnika Poznańska, Poznań, krzysztof.wesolowski@put.poznan.pl

DOI: 10.15199/59.2023.4.88

Streszczenie: W artykule rozważono problem zmniejszenia strat zasobów radiowych i opóźnienia w ich alokacji, gdy zastosowano zasadę scentralizowanego dynamicznego ich przydziału (Dynamic Grant - DG) w transmisji V2X New Radio. Zaproponowano dwie metody wykorzystujące zmienną wielkość podkanałów, aby zredukować straty i opóźnienie. Metody wykorzystują zarówno przylegające jak i nieprzylegające bloki zasobów (RB). Uzyskane zmniejszenie strat zasobów i opóźnienia jest szczególnie widoczne w przypadku zastosowania nieprzylegających bloków zasobów.

Abstract: We investigate the problem of decreasing the waste of resource blocks (RB) and resource selection delay in the Dynamic Grant (DG) centralized resource (re)allocation of the New Radio V2X communication. We propose two methods that use variable subchannel sizes to reduce RB waste and resource selection delay. The algorithms use adjacent/non-adjacent RBs to carry the data packets. The resulting decrease in wasted RBs and average allocation delay with non-adjacent RBs is particularly promising.

Słowa kluczowe: Komunikacja między pojazdami V2X NR, scentralizowany dynamiczny przydział zasobów.

Keywords: V2X NR vehicular communication, centralized dynamic resource allocation.

1. WSTĘP

Komunikacja między pojazdami oraz pojazdami i innymi obiektami (V2X - *Vehicular-to-Everything*) jest jedną z ważnych usług w systemach 5G. Konsorcjum 3GPP w 16. wydaniu standardów znormalizowało zastosowanie komunikacji V2X 5G New Radio [1], [2], w której interfejs radiowy zapewnia komunikację bezpośrednią (SL - *Side-Link*) z otaczającymi dany pojazd obiektami [3].

W paśmie działania systemów V2X NR zbior zasobów realizujących transmisję SL składa się z przylegających do siebie podkanałów [4]. Wielkość każdego podkanału jest ustalona i sformowana z N rozmieszczonych obok siebie bloków RB. W systemach V2X NR V2X możliwe są podkanały składające się z $N = 10, 12, 15, 20, 25, 50, 75$, lub 100 bloków zasobów RB [5].

W dziedzinie czasu szczeliny czasowe wykorzystywane do transmisji bezpośredniej SL są określone w powtarzanych okresowo binarnych mapach [6]. Transmisja SL w 5G NR V2X jest możliwa w trybie 1 ze scentralizowaną alokacją zasobów lub w trybie 2 z ich autonomiczną alokacją. W trybie 1 infrastruktura komórkowa (tj. stacje bazowe gNB) zarządza wyborem zasobów w ramach komunikacji SL wykorzystując techniki DG (*Dynamic Grant*) lub CG (*Configure Grant*). Terminale (UE) działające w technice DG muszą wysyłać żądanie przydziału zasobów (SR - *Scheduling Request*) do stacji bazowej gNB przy transmisji każdego bloku transportowego (TB). Z kolei, metoda alokacji zasobów CG przyznaje terminalowi UE więcej niż jeden interwal na transmisję bloku transportowego, lecz nie więcej niż trzy. Aby realizować transmisję danych sterujących i bloku transportowego stosowane są dwa kanały fizyczne sąsiadujące ze sobą lub nie na osi częstotliwości.

Jedną z głównych wad alokacji zasobów w systemach V2X są straty wynikające z niewykorzystanych bloków zasobów, gdy podkanały mają stały wymiar i są zbudowane z bloków zasobów umieszczonych obok siebie, co jest szczególnie widoczne, gdy liczba bloków RB w podkanałach jest duża. Wymiar podkanałów jest zazwyczaj mniejszy niż niezbędny do transmisji pojedynczego pakietu, należy więc zastosować kilka podkanałów.

W niniejszym artykule proponujemy dwie nowe metody alokacji zasobów dla systemów V2X NR stosujących zasadę dynamicznego przydziału DG. Metody te pozwalają na zastosowanie zmiennej wielkości podkanałów z przylegającymi lub nieprzylegającymi do siebie blokami zasobow RB. Dzięki nim zredukowane mogą być straty wynikające z niewykorzystania zarezerowanych zasobów i czas potrzebny do wyboru odpowiedniego podkanału.

2. SFORMUŁOWANIE PROBLEMU

Zgodnie z ideą części pasma (*Bandwidth Part*) [3], w celu realizacji komunikacji 5G V2X SL [7] 3GPP przydzieliło jedno podpasmo. W konsekwencji wszystkie terminale V2X stosują te same parametry transmisji OFDM w komunikacji SL. Odstęp między podnośnymi OFDM (SCS - *Subcarrier Spacing*) wynosi więc 15 kHz, zaś każdy blok zasobów RB składa się z 12 podnośnych. Podkanały składają się z kolei z grup bloków RB.

W naszych badaniach założono, że każdy terminal transmituje pakiety z okresem T_{step}, zaś stacja bazowa gNB wybiera podkanały dla pojazdów znajdujących się w jej zasięgu radiowym. Liczba bloków RB określających podkanał może być wyznaczona ze wzoru:

$$RB_{SCH} = RB_{TB} + \alpha \cdot RB_{PSCCH} + RB_{free} \quad (1)$$

w którym RB_{TB} oraz RB_{PSCCH} oznaczają odpowiednio liczbę bloków RB użytych do transmisji bloku transportowego TB (*Transport Block*) i liczbę bloków RB przenoszących informacje sterujące (2 bloki RB) fizycznego kanału sterującego transmisją bezpośrednią PSCCH (*Physical Sidelink Control Channel*). Symbol RB_{free} oznacza liczbę nieużytych bloków RB w podkanale. Współczynnik α jest równy 1 lub 0 odpowiednio w przypadku zastosowania przylegających lub nieprzylegających RB_{TB} bloków i RB_{PSCCH} bloków RB. Liczba SCH_{slot} podkanałów o stałym wymiarze w szczelinie czasowej może zostać wyznaczona z zależności:

$$SCH_{slot} = \left\lfloor \frac{RB_{slot}}{RB_{SCH}} \right\rfloor \tag{2}$$

w której RB_{slot} jest liczbą bloków RB w każdej szczelinie czasowej. Podstawiając wzór (1) do wzoru (2) i oznaczając jako $RB_r = RB_{TB} + RB_{PSCCH}$ liczbę bloków RB zawierających dane, SCH_{slot} możemy zapisać jako:

$$SCH_{slot} = \left\lfloor \frac{RB_{slot}}{RB_r + RB_{free}} \right\rfloor \tag{3}$$

Liczba Q szczelin czasowych między dwoma kolejnymi rozgłoszeniami pakietów przez ten sam terminal UE wynosi $Q = T_{step}/T_{slot}$, gdzie T_{slot} to czas trwania szczeliny czasowej. Maksymalną liczbę terminali, które rozgłaszają wiadomości o swoim stanie w danym obszarze możemy otrzymać jako równą $V = Q \cdot SCH_{slot}$.

Część bloków zasobów może być zmarnowana, gdy liczba pozostałych bloków w szczelinie jest mniejsza niż zakładany wymiar podkanału. Liczba RB_{slot} bloków RB w założonym paśmie transmisji (np. 10 MHz) może być określona jako:

$$RB_{slot} = SCH_{slot}(RB_r + RB_{free}) + B_{slot} \tag{4}$$
$$\text{pod warunkiem: } 0 \leq B_{slot} < RB_{SCH},$$

gdzie B_{slot} jest liczbą zmarnowanych bloków RB w pojedynczej szczelinie czasowej. B_{slot} może zostać wyznaczony z użyciem (2) za pomocą następującego wzoru:

$$B_{slot} = RB_{slot} - SCH_{slot} RS_{SCH} \left\lfloor \frac{RB_{slot}}{SCH_{slot} RS_{SCH}} \right\rfloor \tag{5}$$

Optymalizacyjny problem mający na celu zwiększenie wykorzystania zasobów przez minimalizację liczby zmarnowanych bloków RB w Q szczelinach czasowych może być sformułowany w następujący sposób:

$$\min_{B_{slot}, RB_{free}} \sum_{j=1}^{Q} \Upsilon_{slot,j} \tag{6}$$

gdzie $\Upsilon_{slot,j}$ jest liczbą zmarnowanych bloków RB w j-tej szczelinie czasowej $(j = 1, ..., Q)$.

Według (5) pełna liczba zmarnowanych bloków RB w pojedynczej szczelinie czasowej wynosi $\Upsilon_{slot,j}$ i dla Q kolejnych szczelin możemy wyliczyć łączną ich liczbę Υ_Q w następujący sposób zależnie od jednego z trzech przypadków:

Przypadek 1: Jeśli oznaczymy RB_r jako wymiar każdego pakietu, to RB_r oraz wielkość podknału są stałe dla wszystkich wybranych podkanałów, jak zaproponowało 3GPP [5]. Wtedy B_{slot} jest również stałe, więc:

$$\Upsilon_{slot,j} = B_{slot} + SCH_{slot} RB_{free}, \quad (j = 1, ..., Q) \tag{7}$$

oraz

$$\Upsilon_Q = Q\Big(B_{slot} + SCH_{slot} RB_{free}\Big) \tag{8}$$

Z powodu ustalonej liczby bloków RB_r w każdym podkanale liczba zmarnowanych bloków RB jest również stała.

Drugim rozważanym problemem jest opóźnienie w alokacji zasobów i ich transmisji. Opóźnienie transmisji τ może wystąpić, gdy czas pomiędzy transmisją dwóch kolejnych pakietów jest dłuższy niż ustalony przez standard 3GPP. Może ono być opisane w następujący sposób:

$$\tau = \begin{cases} 0 & t_1 = (t_0 + T_{step}) \\ t_{del} & t_1 \neq (t_0 + T_{step}) \end{cases}, \quad t_{del} = t_1 - (t_0 + T_{step}) \tag{9}$$

gdzie t_0 oraz t_1 są odpowiednio chwilami użycia szczelin czasowych ostatniego transmitowanego pakietu i dla nowo wybranego podkanału.

3. PRZYLEGAJĄCE I NIEPRZYLEGAJĄCE BLOKI ZASOBÓW W PODKANALE

Bardziej elastyczne wykorzystanie zasobów przez terminale UE może skutkować większą liczbą pojazdów, które mogą transmitować wiadomości w zadanym obszarze. Stacja bazowa gNB może lepiej wykorzystywać zasoby radiowe przez zastosowanie zmiennej liczby N bloków RB w każdym podknale, aby lepiej dopasować ją do liczby RB_r. Tak więc straty bloków RB mogą zostać zredukowane, jak to ma mmiejsce w następującym przypadku:

Przypadek 2: Liczba RB_r jest zmienna w każdym pakiecie z górną granicą równą ustalonemu wymiarowi danego podkanału dla stałych wymiarów podkanałów. Wtedy liczba B_{slot} jest stała we wszystkich szczelinach czasowych, a w konsekwencji:

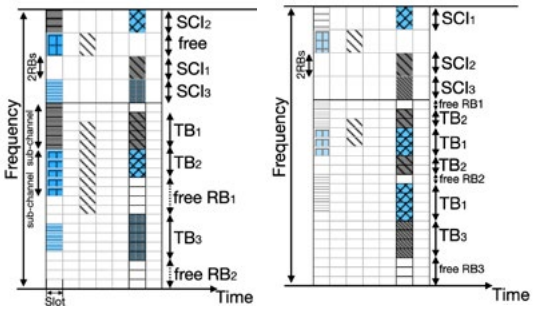
$$\Upsilon_{slot,j} = B_{slot} + \sum_{i=1}^{SCH_{slot}} RB_{free,i}, \quad (j = 1, ..., Q) \tag{10}$$

$$\Upsilon_Q = \sum_{j=1}^{Q} \Upsilon_{slot,j} = \sum_{j=1}^{Q} \left(B_{slot} + \sum_{i=1}^{SCH_{slot}} RB_{free,i,j} \right) \tag{11}$$

W przypadku bloków zasobów przylegających do siebie w podkanale stacja gNB sprawdzi ostatnią szczelinę czasową użytą do transnmisji przez dany terminal. Jeśli wszystkie lub część zmarnowanych zasobów są blokami przylegającymi do siebie i ich liczba jest nie mniejsza niż RB_r, to stacja gNB poinformuje pojazd, aby użył on te bloki RB jako następny podkanał (por. rys. 1a).

W przypadku przeciwnym, gdy ostatnio wykorzystana szczelina czasowa nie zawierała wystarczającej liczby wolnych zasobów (tj. liczba przylegających bloków RB jest mniejsza niż RB_r), wtedy stacja gNB będzie poszukiwać następną szczelinę czasową, aż znajdzie wystarczające miejsce a wtedy poinformuje o tym pojazd. Liczba tych sprawdzeń implikuje wielkość opóźnienia w przydziale zasobów.

Wada tej metody ujawnia się wtedy, gdy nowa wiadomość wymaga większej liczby zasobów niż liczba istniejących wolnych przylegających bloków RB. Aby wyeliminować tę niedogodność, proponujemy przydział nieprzylegających do siebie bloków RB w podkanale, co wspomaga poprzednią metodę przydziału zasobów.



(a) Bloki przylegające (b) Bloki nieprzylegające

Rysunek 1: Konfiguracje bloków zasobów RB

Tak więc stacja gNB mogłaby stosować odseparowane od siebie o zmiennej liczbie bloki RB w danej szczelinie czasowej (por. rys. 1b). Stacja gNB powinna sprawdzić, ile bloków RB jest potrzebnych do transmisji RB_r. Co więcej, stacja gNB powinna określić liczbę zmarnowanych bloków RB w ostatnio używanej szczelinie czasowej. Jeśli Υ_{slot} jest większe lub równe RB_r, to gNB poinformuje pojazd, aby zajął on te bloki RB w następnym podkanale w postaci nieprzylegających bloków RB w aktualnej szczelinie czasowej. W przeciwnym przypadku jeśli Υ_{slot} jest mniejsze niż RB_r, stacja gNB weźmie pod uwagę następną szczeinę czasową, albo kolejne. W takim przypadku liczba zmarnowanych bloków RB może być wyliczona w poniższy sposób.

Przypadek 3: RB_r składa się z bloków RB rozmieszczonych w tej samej szczelinie czasowej w sposób nieprzylegający do siebie. W tym przypadku wartość RB_{free} jest równa zeru a liczba wolnych bloków RB w tej samej szczelinie wynosi B_{slot}. (B_{slot} jest mniejsze niż najmniejsza liczba L_{packet} bloków RB wymaganych do przesłania pakietu), ktora jest zmienna dla wszystkich podkanałów:

$$\Upsilon_{slot,j} = B_{slot,j}, (j = 1, ..., Q), \quad \Upsilon_Q = \sum_{j=1}^{Q} \Upsilon_{slot,j} \quad (12)$$

Zwróćmy uwagę, że wybrane bloki zasobów muszą być wskazane w około jednym bajcie dodatkowej informacji zawartej w bloku sterującym SCI (*Sidelink Control Information*), który jest rozgłaszany w części sterującej transmitowanych pakietów. Część ta jest rozgłaszana z użyciem kanału PSCCH (*Physical Sidelink Control Channel*).

4. SYMULACJE SYSTEMU

W wyniku symulacji wyznaczono wielkość niewykorzystanych/zmarnowanych zasobów w przypadkach 2 oraz 3. Dla wspomnianych przypadków wyliczano Υ_Q (równania (11) i (12)). Wyznaczano również opóźnienie w alokacji zasobów dla obu proponowanych metod. W tym celu stosowano uśrednianie wyników wartości Υ_Q względem eksperymentów Monte Carlo.

Zaproponowane metody alokacji były testowane dla różnych gęstości ruchu pojazdów. Główne parametry symulacji zostały przedstawione w tabeli 1. W pierwszym przebiegu symulacyjnym zakładano, że zasoby używane przez wszystkie pojazdy zostały umieszczone na scentralizowanej siatce zasobów i miały zarezerwowane własne podkanały. Każdy pojazd w trakcie symulacji otrzymywał dynamicznie nowe zasoby co T_{step} (100 ms). Krok symulacji odpowiadał odstępowi T_{step}.

Tabela 1: Główne parametry symulacji

Parametry	*Wartości*
Szerokość pasma	10 MHz
Numer MCS	zmienny
Liczba podkanałów w szczelinie	zmienna
Min. wielkość pakietu L_{packet}	6 RB
Maks. wymiary podkanałów	10, 12 i 15 RB
Liczba pojazdów	100 do 320
Długość szczeliny czasowej	1 ms
Okres wiadomości CAM T_{step}	100 ms
Liczba alokacji zasobów	100000
Odstęp między podnośnymi (SCS)	15kHz
Liczba bloków RB w szczelinie	50
Liczba używanych bloków RB w szczelinie	45

W przypadkach 2 oraz 3 liczba RB_r była generowana losowo w przedziale pomiędzy dolną i górną granicą wynikającą z wielkości N bloków RB wymiaru podkanału (tj. dolną granicą jest 6 bloków RB zaś górną jest 15 bloków).

Symulacje koncentrowały się na wyborze zasobów radiowych z użyciem jednej z proponowanych metod dla struktury komórkowej obliczając liczbę zmarnowanych zasobów w zakładanej szczelinie czasowej ignorując efekty związane z mobilnością i zmiennością kanałów.

5. WYNIKI SYMULACJI

Wyniki symulacji uzyskano dla dwóch proponowanych metod alokacji zasobów. Na rys. 2, 3 i 4 zaprezentowano jakość działania metod scentralizowanej alokacji zasobów dla stacji bazowych gNB określając liczbę zmarnowanych zasobów oraz opóźnienie w ich alokacji w przypadku stosowania przylegających i nieprzylegających bloków RBs transportujących pakiety.

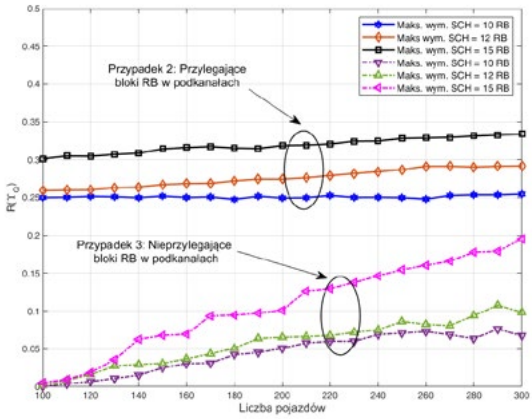
Gdy wymiar podkanału jest stały i duży, liczba zmarnowanych zasobów będzie zwiększona. W przypadku 2 zilustrowanym na rys. 2 pokazano większą stratę bloków RB szczególnie wtedy, gdy podkanał składa się z 15 przylegających do siebie bloków RB. Strata ta zmniejsza się ze zmniejszaniem się wymiaru podkanału, co pokazano, gdy górną granicą wymiaru podkanału jest odpowiedni 12 i 10 bloków RB.

Zwróćmy uwagę, że strata bloków RB w przypadku 2 zmienia się mniej, gdy liczba pojazdów w sieci rośnie niż w przypadku 3, ponieważ liczba bloków RB jest stała a więc storunek (9) liczby zmarnowanych bloków RB do użytych podkanałów jest taki sam.

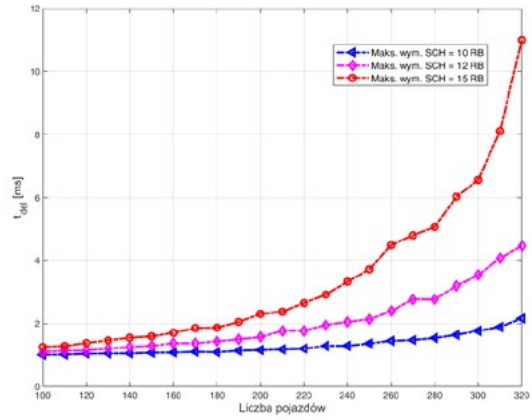
Na rys. 3 i 4 przedstawiono opóźnienie w alokacji zasobów odpowiednio dla przypadków 2 i 3. Jest oczywiste, że gdy liczba pojazdów w sieci jest niewielka, opóźnienie w alokacji zasobów jest niskie ponieważ niskie jest obciążenie kanału. Jednak opóźnienie alokacji zasobów zwiększa się, gdy obciążenie kanału wzrasta. Zwiększa się ono proporcjnalnie ze znalezieniem w pełni zajętych zasobów kanału w pożądanej szczelinie czasowej (tj. gdy liczba zmarnowanych bloków RB jest niższa niż liczba żądanych bloków niezbędnych do transmisji nowego pakietu.

Zmienna liczba nieprzylegających do siebie bloków RB stosowanych do transmisji pakietów pokazana dla przypadku 3 powoduje mniejszą liczbe zmarnowanych bloków RB. Spadek ten wynika z lepszego wykorzystania pasma kanału. Można uznać, że to obiecujący rezultat, bowiem RB_r bloków zasobów stosowanych przez pakiet można umieścić w miejscach pomiędzy blokami zasobów użytych przez inne pakiety. Jednak można zaobserwować pewną liczbę zmarnowanych pakietów, gdy RB_r rośnie.

Dlatego straty zasobów stają się szczególnie widoczne, gdy liczba pojazdów w sieci wzrasta i wzrasta obciążenie kanału. Zerową stratę bloków RB można zaobserwować, gdy liczba pojazdow jest równa 100 i stopniowo wzrasta wraz z RB_r do osiągnięcia wartości ułamka zmarnowanych zasobów (9) równej 0.2, gdy liczba pojazdów wynosi 300 i gdy górną granicą użytych bloków RB jest 15.



Rysunek 2: Ułamek zmarnowanych bloków RB dla alokacji pakietów w przypadkach 2 1 3.

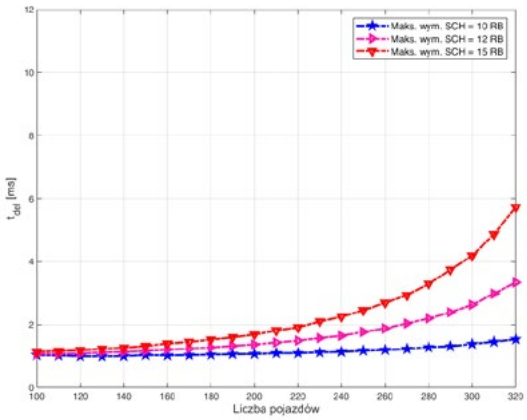


Rysunek 3: Opóźnienie alokacji zaobów gdy zastosowano metodę przydziału przylegających bloków RB.

Na rys. 3 można zauważyć, że największe opóźnienie występuje wtedy, gdy górna granica RB_r wynosi 15 bloków RB. Wtedy liczba wymaganych przylegających bloków zasobów jest wysoka i znalezienie wymaganej liczby zasobów w paśmie kanału nie jest osiągalne.

Najniższe opóźnienie wynosi 1 ms, gdy liczba pojazdów/stacji UE jest niska. Ono stopniowo wzrasta wraz z rosnącą liczbą terminali. Górna granica opóźnienia w alokacji żądanych bloków RB osiąga wartość około 11 ms dla górnej granicy 15 bloków RB, podczas gdy opóźnienie dochodzi odpowiednio do 4.3 ms oraz 2.1 ms dla górnej granicy 12 i 10 bloków RB.

Z rys. 4 można wywnioskować, że jakość działania w przypadku zastosowania przydziału nieprzylegających do siebie bloków RB w podkanale jest obiecująca w porównaniu do metody alokacji z przylegającymi blokami RB, co pokazano na rys. 3. Jak widać, w tym przypadku obserwuje się opóźnienie w alokacji zasobów około dwa razy mniejsze niż na rys. 3.



Rysunek 4: Opóźnienie alokacji zasobów gdy zastosowano metodę z nieprzylegającymi blokami RB.

6. WNIOSKI

W niniejszej pracy zaproponowano dwie metody zmniejszenia strat bloków RB oraz opóźnienia w dynamicznej alokacji zasobów w podkanałach w transmisji bezpośredniej (SL) systemu NR V2X w trybie 1 w porównaniu z typową konfiguracją, gdy podkanały dysponują stałą liczbą bloków zasobów RB przylegających do siebie. W zaproponowanych metodach sprawdzono zastosowanie przylegających do siebie i nieprzylegających bloków RB i ich wpływ na zmniejszenie strat zasobów i opóźnienia w ich alokacji. Okazuje się, że umożliwienie przydziału nieprzylegających bloków prowadzi do znaczącego ograniczenia strat zasobów i opóźnienia w ich alokacji.

Praca finansowana z projektu 0312/SBAD/8162.

LITERATURA

[1] 3GPP, RP-170837, "New SI Proposal: Study on Evaluation Methodology of New V2X Use Cases for LTE and NR," 3GPP TSG RAN Meeting 75, Dubrovnik, Croatia, March 2017.

[2] M. H. C. Garcia, A. Molina-Galan, M. Boban, J. Gozalvez, B. Coll-Perales, T. Şahin, and A. Kousaridas "A Tutorial on 5G NR V2X Communications," IEEE Communications Surveys and Tutorials, vol. 23, no. 3, Third Quarter 2021, doi: 10.1109/COMST.2021.305701. pp. 1972-2026.

[3] 3GPP TS 38.211, "5G; NR; Physical Channels and Modulation", Rel-16, V16.0.0, 3GPP, Tech. Spec., Jul. 2020.

[4] 3GPP TS 38.214, "TSG RAN; NR; Physical Layer Procedures for Data," Rel-15, v16.4.0, Jan. 2021.

[5] 3GPP TS 38.202, "TSG RAN; NR; Services Provided by the Physical Layer," Rel-15, v16.2.0, Sep. 2020.

[6] 3GPP TS 38.331, "TSG RAN; NR; Radio Resource Control (RRC); Protocol specification," Rel-15, v16.2.0, Sep. 2020.

[7] 3GPP Tech. Rep., TR 37.985, "Overall Description of Radio Access Network (RAN) Aspects for Vehicle-to-Everything (V2X) Based on LTE and NR," Release 16, v16.0.0, Sept. 2020.

Automatyzowanie procesów zarządzania zmianami w sieciach

Automating change management processes in networks

Wiktor Tekiela, Jerzy Domżał

Akademia Górniczo-Hutnicza im. Stanisława
Staszica w Krakowie,
wiktor.tekiela@student.agh.edu.com,
jerzy.domzal@agh.edu.pl

DOI: 10.15199/59.2023.4.89

Streszczenie: Automatyzacja od wielu lat jest powszechnym zjawiskiem w sieciach IP. Jest to rozwiązanie szczególnie popularne w rozległych oraz większych sieciach lokalnych. Dzięki rozwojowi tego sektora, administratorzy mogą uchronić swoje sieci przed błędami, w których główną rolę odgrywa czynnik ludzki. Szczególny rozwój odnotował sektor odpowiedzialny za wdrażanie nowych rozwiązań w sieci dzięki podejściu "infrastruktura jako kod" oraz dzięki wprowadzeniu i upowszechnieniu protokołów takich jak RESTCONF.

Abstract: Automation in networks has been a common phenomenon in wide area networks and larger local networks for several years. Thanks to the development of this sector, administrators can protect their networks from errors in which the human factor plays a major role. Particular development was recorded in the sector responsible for implementing new solutions in the network thanks to the approach of infrastructure as code and thanks to the introduction and dissemination of protocols such as RESTCONF.

Słowa kluczowe: Automatyzacja, DevOps, Infrastruktura jako kod, Zarządzanie zmianami.

Keywords: Automation, Change Management, DevOps, Infrastructure as a code.

1. WSTĘP

Automatyzowanie wdrażania zmian w sieci przeszło na przestrzeni lat ewolucję. Pierwotnie używano w tym celu skryptów przygotowywanych w BASH oraz skryptów wykorzystujących protokół SNMP. Rozwiązania te były dużo prostsze w porównaniu ze współcześnie stosowanymi, gdyż wymagały one jedynie uruchomienia lub zakolejkowania przy użyciu linuxowego narzędzia CRON lub podobnego. Punktem, który przyspieszył rozwój tego procesu było opracowanie protokołu NETCONF, który uprościł rekonfigurację urządzeń sieciowych i pozwolił na tworzenie bardziej generycznych skryptów. W dzisiejszych czasach tworzenie infrastruktury sieciowej opiera się nie tylko na fizycznych urządzeniach, ale również na usługach chmurowych. Podejście "infrastruktura jako kod" znacząco uprościło integrację chmury z urządzeniami sieciowymi i otworzyło możliwość automatyzacji, a w niektórych przypadkach standaryzacji, procesu wdrażania zmian w infrastrukturze. Automatyzacja tego procesu przebiega w podobny sposób jak automatyzacja wdrażania zmian w oprogramowaniu, gdzie kod przed wdrożeniem jest badany pod kątem jego jakości, bezpieczeństwa, ale również dzięki testom pod kątem możliwych błędów logicznych. Większą część automatyzacji wprowadzania zmian w sieci można przenieść z procesu wdrażania zmian w kodzie [2]. Warto zaznaczyć, że implementacje automatyzujące [5] mogą się znacząco różnić między sobą, jednak w większości z nich powinniśmy być w stanie wyróżnić cztery główne etapy:

1. walidacja składni kodu,
2. przeprowadzanie testów jednostkowych,
3. walidacja wydajności,
4. wdrożenie zmiany.

Każdy z tych etapów został zaimplementowany w systemie proponowanym w tym artykule, który porównano z popularnym rozwiązaniem powszechnie stosowanym w automatyzowaniu sieci.

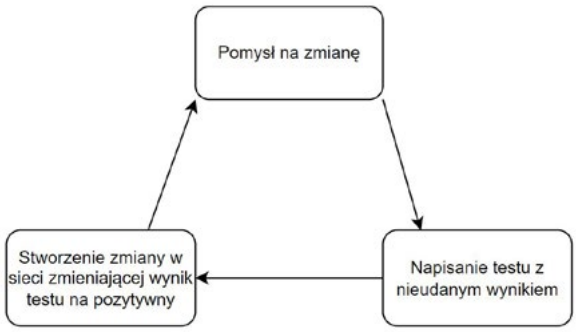
2. WALIDACJA SKŁADNI KODU

Walidacja składni kodu to pierwszy etap procesu wdrażania zmian. W tym miejscu administrator sieci definiuje topologię sieci w formie kodu przy użyciu plików YAML, HCL lub JSON. System automatyzujący powinien sprawdzić poprawność składni danego formatu oraz czy formaty zdefiniowanych urządzeń są właściwe. Oprócz tego, w tym etapie możliwa jest implementacja sprawdzania infrastruktury i wyszukiwania w niej podatności na ataki lub przeciążenia na bazie wcześniej zdefiniowanych sygnatur. Narzędzia dostępne na rynku, takie jak Terraform [7] czy Ansible [8], posiadają wbudowane skanery podatności oraz kompilatory, które potrafią wykryć błąd w schemacie infrastruktury. Na potrzeby badań zastosowaliśmy Ansible oraz autorski system [6].

3. PRZEPROWADZANIE TESTÓW JEDNOSTKOWYCH

Kolejnym etapem po walidacji schematu infrastruktury jest automatyczne przeprowadzanie testów jednostkowych zmian, które administrator sieci chce wprowadzić. W idealnym scenariuszu zmiany te powinny być testowane na jednoznacznie odwzorowanej kopii sieci. Podejście takie jest możliwe w infrastrukturze w pełni opartej na chmurze. Jednak jeśli mamy do czynienia z siecią hybrydową lub składającą się w pełni z urządzeń fizycznych, to zapewnienie kopii tych urządzeń oraz ich utrzymanie jest

zbyt kosztowne. Alternatywą dla kopii sieci jest testowanie zmian w czasie kiedy są one wyłączone z użytku lub kiedy z sieci korzysta niewielu użytkowników. Bezpieczniejszym rozwiązaniem, jednak nie zawsze możliwym do zrealizowania, jest wirtualizacja fizycznych urządzeń w sieci. Wielu producentów dostarcza gotowe obrazy, które mogą zostać wykorzystane w Dockerze lub w środowisku wirtualizacyjnym jak GNS3. Testy jednostkowe, podobnie jak te wykorzystywane do testowania oprogramowania, muszą zostać wcześniej przygotowane przez administratora bądź DevOps'a. Testy te powinny pokrywać jak największą część infrastruktury i jej kluczowych funkcjonalności. Jednym z podejść do wdrażania zmian w sieci lub nawet jej budowania może być podejście Test Driven Development, którego schemat został zaprezentowany na rysunku numer 1. Podejście to jest wykorzystywane między innymi do wytwarzania oprogramowania, ale można go użyć z powodzeniem również w procesie wdrażania zmian w sieci. Rozwiązanie to zakłada utworzenie testów jednostkowych, w których definiujemy testy funkcji, których oczekujemy od wdrażanej sieci. Do pisania samych testów jednostkowych sieci możemy skorzystać z wielu dedykowanych do tego bibliotek, takich jak na przykład AETest z języka Python.



Rysunek 1: Podejście Test Driven Development w rozwoju sieci [1].

4. WALIDACJA WYDAJNOŚCI

Ostatnim etapem sprawdzającym zmiany przygotowane do wdrożenia jest walidacja wydajności. Celem tego etapu jest sprawdzenie, czy sieć zachowuje się odpowiednio w różnych scenariuszach testowych, w ramach których jest poddawana obciążeniom. Podobnie jak w etapie przeprowadzania testów jednostkowych administrator musi mierzyć się z przygotowaniem środowiska testowego. Etap ten nie jest generyczny w porównaniu do testów jednostkowych, które w każdej implementacji procesu wdrażania zmian wyglądają podobnie. Często nie pojawia się on w implementacji jako oddzielna część. W celu uproszczenia niejednokrotnie stosowany jest prosty generator ruchu w etapie testów jednostkowych. Walidacje wydajności w sieci mogą występować w kilku rodzajach.

4.1 Testy obciążeniowe

Są podstawowym typem testu wydajności. Wykorzystują one jednostajne obciążenie systemu i sprawdzają, jak zmieniają się parametry sieci. Rodzajem tego typu testów są testy wytrzymałościowe.

4.2 Testy przeciążeniowe

Testy przeciążeniowe wykorzystują ruch znacznie większy niż pojemność sieci. Testy te sprawdzają, czy pomimo tak wysokiego obciążenia, sieć zapewnia odpowiednie pasmo dla ruchu z wyższym priorytetem i ruchu sygnalizacyjnego. Jednym z rodzajów takich testów są testy skokowe.

Przy implementacji tego etapu należy dobrać odpowiednie testy do wymagań sieci. Implementacja wszystkich testów może być kosztowna oraz może znacząco przedłużać czas całego zautomatyzowanego procesu, często niepotrzebnie.

5. WDROŻENIE ZMIANY

Po pomyślnym wyniku wszystkich etapów następuje wdrożenie zmian do właściwej sieci. Przed samym wdrożeniem zautomatyzowany proces powinien utworzyć obraz każdego urządzenia w sieci. Obrazy te w połączeniu z plikami konfiguracyjnymi z infrastruktury jako kodu pozwalają na przechowywanie stanu sieci w systemie kontroli wersji takim jak GIT.

6. IMPLEMENTACJA KONTROLERA AUTOMATYZUJĄCEGO WDRAŻANIE ZMIAN W SIECI

Rekonfiguracja urządzeń może odbywać się przy użyciu różnych podejść [4]. Pierwszym podejściem jest podejście scentralizowane, w ramach którego możemy wyróżnić główny kontroler mający możliwość połączenia się z każdym urządzeniem w sieci. Drugie podejście "device-by-device" jest podejściem klasycznym, w którym kontroler jest przemieszczany i oddzielnie podłączany do każdego z urządzeń. Podejście device-by-device przez swoją specyfikę posiada mniejsze możliwości automatyzacji, przez to, że istnieje wiele różnych kontrolerów. Wciąż jednak etapy procesu wdrażania, w tym walidacje, mogą zostać zautomatyzowane i nie wymagają od administratora repetytywnego wykonywania komend na każdym urządzeniu. Podejście to odnajduje zastosowanie w dużych sieciach, gdzie scentralizowany kontroler mógłby posiadać problemy z wydajnością. Scentralizowany kontroler, przez to, że ma możliwość komunikacji z każdym urządzeniem w sieci, posiada jednak większe możliwości. Przede wszystkim testy wydajnościowe w takim kontrolerze nie są przeprowadzane z poziomu jednego urządzenia a z poziomu całej sieci. Podejście z użyciem scentralizowanego kontrolera pozwala na dokładniejszą walidację oraz na większy poziom automatyzacji całego procesu.

7. DOSTĘPNE NARZĘDZIA

7.1 Ansible

Najpopularniejszym narzędziem do automatyzacji procesu rekonfiguracji sieci jest Ansible, przejęty w 2015 roku przez RedHat. Ansible posiada szereg funkcjonalności i rozwiązań. Najpopularniejsze z nich opiera się o playbook'i, czyli pliki w formacie YAML, w ramach których użytkownik definiuje zadania do wykonania przez Ansible. Zadania zawierają w swojej definicji akcję oraz urządzenia, na których ta akcja ma zostać wykonana. Ansible reprezentuje podejście scentralizowane, dlatego jesteśmy w

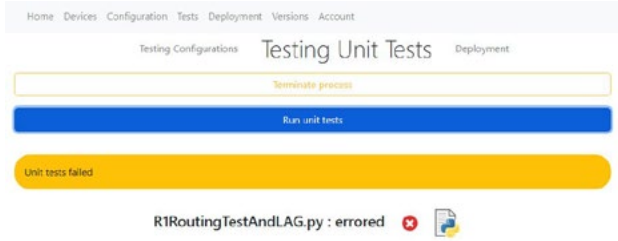
stanie wyróżnić w nim węzeł sterujący, czyli węzeł z zainstalowanym oprogramowaniem Ansible, oraz węzły zarządzalne, którymi są wszystkie urządzenia w sieci lub tylko wybrane, na których proces wdrażania zmiany ma zostać zautomatyzowany.

7.2 Puppet

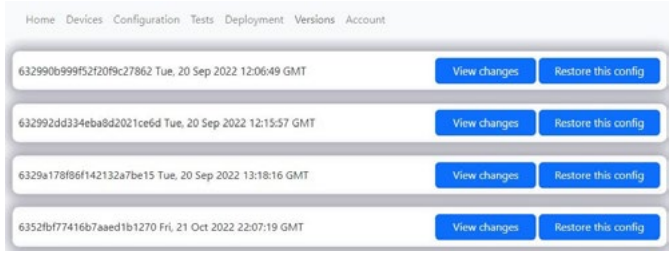
Puppet jest narzędziem, które pozwala zarządzać konfiguracją urządzeń w sposób zautomatyzowany. Oprogramowanie zostało udostępnione w 2005 roku. Zasada działania tego narzędzia znacząco się różni od Ansible, gdyż Puppet wymaga do poprawnego działania agenta na urządzeniach końcowych. Jednak dzięki tej specyfice Puppet jest w stanie monitorować urządzenia w interwałach czasowych i kontrolować czy konfiguracje urządzenia są tożsame z tymi zdefiniowanymi w głównym kontrolerze. Standardowe interwały, w których agent zainstalowany na urządzeniu zarządzalnym sprawdza stan konfiguracji, trwają 30 minut.

7.3 Niestandardowy system z GUI

Najpopularniejsze rozwiązania standardowo nie oferują wszystkich wymienionych wcześniej walidacji [3]. Ansible może zostać skonfigurowany w taki sposób, aby automatyzować wszystkie etapy procesu, jednak wymaga on specjalistycznej wiedzy i poświęcenia dużej ilości czasu, aby dostosować konfigurację do istniejącej sieci. Rozwiązaniem proponowanym w tym artykule jest aplikacja sterująca procesem wdrażania zmian w sieci. Podobnie jak Ansible reprezentuje ona podejście scentralizowane. Korzysta ona z serwera napisanego w języku Python oraz bazy danych NoSQL. Dzięki temu, że aplikacja ta skupia się tylko na automatyzacji procesu wdrażania zmian, to implementuje ona wszystkie wymagane kroki walidacji przedstawione w tym artykule. Aplikacja jest dostępna dla użytkowników w formie webowej. Obecna wersja wspiera urządzenia typu ruter i przełącznik firmy Cisco posiadające system IOS lub IOS-XR. Przy użyciu predefiniowanych formularzy użytkownik może zdefiniować w uproszczonej formie urządzenia ze swojej infrastruktury. Wiele operacji takich jak budowanie infrastruktury w formie kodu jest wykonywane przez aplikację w tle bez udziału użytkownika. Aplikacja posiada generyczny system zarządzania testami jednostkowymi i wydajnościowymi, które są używane w późniejszych walidacjach. Aplikacja obecnie jest w stanie obsłużyć testy napisane przy użyciu biblioteki AEtest stworzonej przez Cisco. Biblioteka ta jest dedykowana do testowania pojedynczych funkcjonalności sieci, ale również umożliwia w prosty sposób przeprowadzenie testów wydajnościowych sieci. Obrazy zdefiniowanych urządzeń, aktualne oraz planowane konfiguracje są przechowywane w bazie danych, co pozwala na zbieranie danych historycznych dla każdego urządzenia. Samo wdrożenie zostało zaimplementowane w formie jednokierunkowego procesu, który składa się z trzech subprocesów: walidacji składni, przeprowadzania testów jednostkowych oraz wydajnościowych, wdrożenia i zapisu w systemie kontroli wersji. Cały proces reprezentuje podejście przeprowadzania walidacji w docelowej sieci. Najbardziej popularne aplikacje takie jak Ansible również standardowo wykorzystują to podejście. Komunikacja z urządzeniem odbywa się poprzez konsolę w sposób programowalny. Pierwszy etap wdrożenia w zautomatyzowany sposób sprawdza poprawność składni konfiguracji przy użyciu systemu operacyjnego docelowego urządzenia oraz połączenia z urządzeniem. Drugi etap jest dostępny dla użytkownika wyłącznie po przejściu pierwszego etapu z wynikiem pozytywnym. Rozpoczęcie drugiego etapu blokuje możliwość wprowadzania zmian w docelowej konfiguracji, co uniemożliwia ingerencję administratora w testy. Etap ten podobnie jak pierwszy jest wykonywany na realnej infrastrukturze i uruchamia on wcześniej zdefiniowane testy w sieci z wdrożoną docelową konfiguracją. Całość jest wykonywana bez ingerencji użytkownika. Po tym etapie sieć jest przywracana do stanu pierwotnego, a użytkownikowi jest prezentowany raport, w którym może on znaleźć potencjalne błędy. Przykładowy wynik testów został zaprezentowany na rysunku numer 2. Ostatni trzeci etap wprowadza zmiany do sieci i zapisuje konfigurację wszystkich urządzeń do specjalnie przygotowanego systemu kontroli wersji, który przechowuje wszystkie dane w bazie NoSQL. Dzięki tej wbudowanej funkcjonalności użytkownik w każdej chwili ma możliwość sprawdzenia historii zmian oraz w razie potrzeby, przywrócenia starszej konfiguracji. Funkcjonalność ta została zaprezentowana na rysunku numer 3.



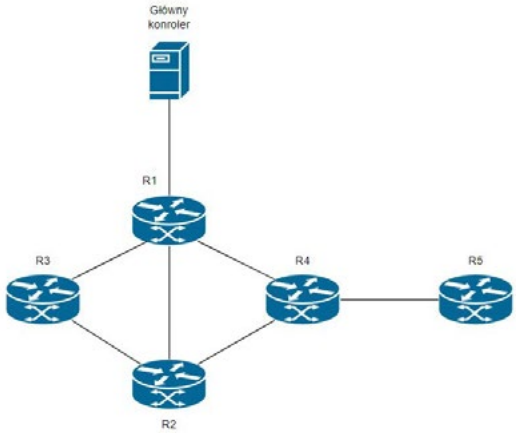
Rysunek 2: Wynik testów jednostkowych w systemie.



Rysunek 3: System kontroli wersji wbudowany w system.

8. PORÓWNANIE WYNIKÓW DWÓCH SYSTEMÓW AUTOMATYZUJĄCYCH

W tym rozdziale przedstawiono porównanie dwóch systemów (wielowątkowego konsolowego oraz jednowątkowego z GUI). W każdym z nich został przeprowadzony proces wdrożenia zmiany, w skład którego wchodzi przetestowanie wcześniej przygotowanej konfiguracji oraz jej wdrożenie. Badanie zostało przeprowadzone z wykorzystaniem topologii przedstawionej na rysunku numer 4. Do badania przygotowano konfigurację protokołu OSPF w formacie YAML oraz w aplikacji GUI. Ponadto przygotowano pięć testów jednostkowych. Trzy z nich testują osiągalność głównego kontrolera, R3 oraz R5 z poziomu każdego urządzenia w sieci. Czwarty test skupia się na sprawdzeniu i zweryfikowaniu tablicy rutingu w ruterze R1. Piąty test weryfikuje liczbę sąsiadów protokołu OSPF w ruterze R4.



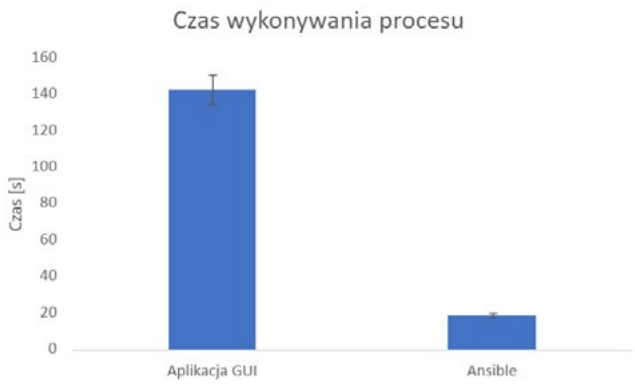
Rysunek 4: Przygotowana topologia sieci wraz z kontrolerem Ansible/Systemu niestandardowego.

8.1 Ansible

Ansible jest aplikacją konsolową wielowątkową przeznaczoną głównie dla systemu Linux. Do przeprowadzenia badania wykorzystany został Ansible w wersji 2.13.7 wraz z pluginem cisco.ios. Jako węzeł sterujący został użyty serwer z systemem Linux Ubuntu. Czas potrzebny serwerowi z Ansible do przeprowadzenia wszystkich operacji wynosił 18,9 +/- 0,19s. Błędy pomiaru wyznaczono z rozkładu t-Studenta na poziomie ufności 0,95. Czas był badany przy użyciu komendy "time" w systemie Linux.

8.2 System niestandardowy

System niestandardowy użyty w badaniu to autorskie rozwiązanie opisane w rozdziale 7.3. Całkowity czas potrzebny na przeprowadzenie walidacji oraz wdrożenie rozwiązania wynosił 143 +/- 8,29 s. Błąd pomiarowy wyznaczono w taki sam sposób jak w przypadku badania Ansible. Czas był mierzony przy użyciu serwera Python i był on mierzony od momentu nadejścia żądania ze strony frontowej do momentu wprowadzenia finalnej konfiguracji we wszystkich urządzeniach. W skład mierzonego czasu wchodziła interakcja użytkownika z interfejsem graficznym oraz komunikacja klient-serwer aplikacji tak, aby zebrane dane były jak najbardziej zbliżone do rzeczywistego scenariusza.



9. PODSUMOWANIE

Porównanie czasu wykonywania tego samego procesu w obu aplikacjach wypada niekorzystnie dla systemu graficznego. Wynik ten jednak nie oddaje całości czasu potrzebnego do wdrożenia zmiany, gdyż porównywany jest tylko czas przeprowadzania operacji. Ansible, jako że jest aplikacją konsolową, wymaga wcześniejszego przygotowania playbook'ów, czyli plików YAML, które definiują zmiany konieczne do wprowadzenia. Jako że przygotowanie tych plików to czasochłonny proces to, Ansible lepiej sprawdzi się w systemach, które służą do diagnostyki i przeprowadzania podobnych zmian okresowo. Aplikacja z GUI lepiej sprawdzi się w topologii, która często jest rekonfigurowana i przebudowywana. Przykładem mogą być nowe sieci, niedawno wdrożone, w których często pojawiają się błędy wymagające szybkiej rekonfiguracji oraz sieci, które są często rozszerzane. Tak drastyczna różnica wynika najprawdopodobniej z tego, że Ansible wykonuje operacje na wielu wątkach w przeciwieństwie do aplikacji z GUI.

LITERATURA

[1] W. Fantom, P. Alcock, B. Simms, C. Rotsos and N. Race, "A NEAT way to test-driven network management," NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium, Budapest, Hungary, 2022, pp. 1-5.

[2] K. Trantzas et al., "An automated CI/CD process for testing and deployment of Network Applications over 5G infrastructure," 2021 IEEE International Mediterranean Conference on Communications and Networking (MeditCom), Athens, Greece, 2021, pp. 156-161.

[3] Singh, C., Gaba, N. S., Kaur, M., Kaur, B. (2019). Comparison of Different CI/CD Tools Integrated with Cloud Platform. 2019 9th International Conference on Cloud Computing, Data Science and Engineering (Confluence).

[4] Chris Jackson, Jason Gooley, Adrian Iliesiu, Ashutosh Malengaonkar. "Cisco Certified DevNet Associate Official Cert Guide". Cisco Press.

[5] S. Mysari and V. Bejgam, "Continuous Integration and Continuous Deployment Pipeline Automation Using Jenkins Ansible," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1-4.

[6] Wiktor Tekiela, "System zarządzania konfiguracją urządzeń w sieci.", https://github.com/abcdrefg/NetworkDeploymentTool, 2023

[7] Terraform, "Introduction to Terraform", https://developer.hashicorp.com/terraform/intro, 2023.

[8] Ansible, "Ansible and Cisco", https://www.ansible.com/integrations/networks/cisco, 2023.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

Practical Models of Interconnections Between Content Delivery Networks and Internet Service Providers

(Praktyczne modele połączeń między operatorami sieci dystrybucji treści i operatorami telekomunikacyjnymi)

Cezary Cłapa[1]; Artur Lasoń[2]; Vinay Kanitkar[3]; Piotr Chołda[4].

[1] Akamai Technologies, Inc; AGH University of Krakow, Poland, cclapa@akamai.com
[2] AGH University of Krakow, Poland, artur.lason@agh.edu.pl
[3] Akamai Technolgoies, Inc, Cambridge, MA, kanitkar@akamai.com
[4] AGH University of Krakow, Poland, piotr.cholda@agh.edu.pl

DOI: 10.15199/59.2023.4.90

Abstract: Content Delivery Networks (CDNs) play a key role in the Internet ecosystem. The continued growth of demand for services delivered online, as well as the increasing bandwidth disparity between its core and the last mile, strengthen their importance. A popular research problem related to CDN is the problem of deploying caching nodes. In this article, we present practical models of interconnection between Internet Service Providers (ISPs) and CDNs. We also enhance, with those models, an existing optimization model that approaches the mentioned problem. While presented interconnection models have a significant impact on the cost structure of the nodes and their scope of connectivity, they are not yet present in the literature.

Streszczenie: Sieci dystrybucji treści CDN (ang. *Content Delivery Networks*) odgrywają kluczową rolę w ekosystemie Internetu. Wzrost zapotrzebowania na usługi dostarczane online i postępująca dysproporcja przepustowości między jego siecią szkieletową, a ostatnią milą wzmacniają ich znaczenie. Popularnym zagadnieniem badawczym związanym z CDN jest problem rozmieszczenia węzłów buforujących. W tym artykule prezentujemy praktyczne modele połączeń między operatorami telekomunikacyjnymi (ang. Internet Service Provider, ISP), a operatorami CDN. Prezentujemy również wzbogacenie o nie przykładowego modelu optymalizacyjnego podejmującego wspominany problem. Modele te mają istotny wpływ na strukturę kosztu węzłów i ich zdolność zaspokajania zapotrzebowania na treści, jednak nie są obecne w literaturze.

Słowa kluczowe: CDN, optymalizacja, ekonomia połączeń międzyoperatorskich, ekonomia CDN

Keywords: Content Delivery Networks, CDN economics, operations research, optimization, interconnectivity

1. INTRODUCTION

A common research problem in the area of Content Delivery Networks (CDNs) is optimizing the deployment of caching nodes. Substantial research has been conducted in this area [5, 9, 10, 7, 11, 4]. In this paper, we discuss different models of interconnections between Internet Service Providers (ISPs) and CDNs. We present how these interconnection models influence the cost of deployment and the scope of connectivity of the caching nodes. These models play a critical role in the deployment decision process due to their diversity in terms of the cost of deployment (and, therefore, serving end-users) and the connectivity scope. Using this insight, we expand a mature deployment optimization model presented in [11].

This paper is composed of six sections: Introduction, Related works, Problem model and notations, Interconnection models (covering Partnership, Private Network Interconnect (PNI), Internet Exchange (IX), Transit models), Example Model Enhancement, and Conclusions.

2. RELATED WORKS

The problem of deploying caching nodes aims at finding an optimal combination of factors influencing the deployment of caching nodes. Those commonly are: location, size, and caching capabilities of the node. Optimization objectives vary between papers, but the metrics that are most often looked at are cost and performance (often called QoS). Performance in this context is understood as the distance between end-users and the nodes [9]. Distance is captured by means of either geographic distance [7] or latency. This approach is based on the observation that the higher the network latency between the node and the user, the higher the chances of unavailable network conditions on the path, e.g. bottlenecks, congestion, or other disturbances [8]. The cost objective is most often captured as the deployment cost, the delivery cost, the energy consumption cost, or the number of servers. In recent years, the research on the subject has focused mostly in the context of Cloud CDNs (CCDNs) — i.e. CDN which are deployed on top of Infrastructure as a Service (IaaS) platforms [11]. Noticeable trends include joint optimization of the deployment factors [10], multi-cloud approaches [3] and dynamic appraches deployments [4, 11]

The problem has been approached with different theoretical models, e.g., Capacited Facility location problem, Minimum K-center, Uncapacitated Facility Location Problem, Splittable Facility Location [7], Uncapacitated K-Median, etc. Regardless of the approach taken, the CDN-ISP interconnection models have not been explicitly captured as on of the deployment factors. In this paper, these interconnection models are presented and their impact on the optimization objectives and constraints of the pro-

blem formulation is discussed. We then focus on a specific example of the optimization model [11] and expand it with insight originating from this paper.

3. PROBLEM MODEL AND NOTATIONS

We follow the authors of [11] in their way of modeling the problem. It aims at deciding on the number of servers to deploy in each location to minimize the deployment cost over time. The relationship between the nodes and the groups of end-users is represented as a bipartite graph. Fig. 1 presents this notation. V is the set of vertices representing the end-users - pictured as 'screens'. L is the set of vertices representing the CDN platform nodes — pictured as nodes. E is the set of edges that represents the assignment of groups of end-users (v) to nodes (l).

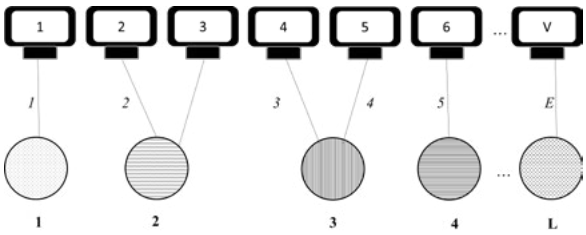


Fig. 1: Bipartite graph representation of the end-users, CDN and their mapping to CDN

Tab. 1: Summary of introduced symbols

Symbol	*Meaning*
p_{hw}	Price of a server hardware
p^l_{colo}	Price of co-locating and powering a server at facility l
p^l_{port}	Per-port fee in IX at facility l
p^l_{xcon}	Per-port fee of provisioning a cross-connect and a router port at facility l
p^l_{bw}	Per-mbps cost of transit service at facility l
V_l	Set of end-users accessible from location l
V_{isp}	Set of end-users accessible though interconnection with an ISP
V_{ix}	Set of end-users accessible though an IX
L_{pni}	Set of facilities where it is possible to connect with an ISP via PNI
L_{pship}	Set of facilities where it is possible to connect with an ISP in a partnership model
L_{ix}	Set of facilities where it is possible to connect with an ISP via IX
L_{transit}	Set of facilities where it is possible to connect with an ISP via Transit
x^l_{min}	Minimal number of servers in a facility necessary in a facility to provide enough disc space to achieve expected caching ratio
T_s	Server's maximal computational ability to deliver content expressed in Gbps

Tab. 1 summarizes the symbols introduced in this paper. These symbols are used to enhance the model presented in [11]. All remaining symbols are inherited from the cited paper. There are three main groups of symbols introduced: 1) those related to different price aspects of interconnected models denoted with p; 2) those related to the scope of connectivity denoted with V; 3) those related to interconnection models available in different facilities denoted with L. T_s denotes computational ability to deliver content of a sever expressed in Gigabits per-second (Gbps).

4. INTERCONNECTION MODELS

In this section, we describe each of the CDN-ISP interconnection models and their impact on the cost component of the interconnection from the perspective of the CDN operator. This insight affects the real-life deployment models.

4.1 Partnership model

A common model of interconnection between ISPs and CDNs is the partnership model. The major operators of CDN networks have their partnership programs, e.g., Google Global Cache, Facebook Network Appliance, Open Connect (Netflix), Akamai Accelerated Network Partner, etc. [1, 2, 6].

In this model, the CDN operator deploys a dedicated node in the ISP network. The CDN provider provides the technology and covers the cost of hardware and remote node operations, including spare parts. The ISP provides colocation, power, connectivity, and on-site support for the node. Usually, the node can only serve end-users of the ISP. The ISP benefits financially from the partnership through reduced expenditures on IP transit services. This is the result of caching: content is downloaded via transit once, cached in the dedicated node of the ISP, and served multiple times from within the network. The CDN benefits financially from the partnership through reduced expenditures on colocation, power, and transit. Both parties benefit from the lower distance between end-users and the caching nodes, as well as reduced dependence on the throughput of the core of the Internet.

This model might seem generally more cost-effective than other models due to fewer cost components. However, there is a minimum number of servers that a caching node needs (to have enough disk space) to achieve the expected caching ratio. Therefore, if that demand that can be served from the node is not sufficient, the cost of deploying that node may not be justified.

4.2 Internet Exchange (IX) Interconnection

In the Internet Exchange interconnection model, the ISP and CDN connect through the infrastructure of an IX. Depending on the financial model of the IX, both the ISP and the CDN might incur the cost of interconnection. Commonly, it is the cost of a port, cross-connects, and deploying a router in the facility operated by the IX, i.e., colocation and power. This model is a popular type of interconnection among smaller ISPs who would like to reach the CDN (and other ISP) directly, but the levels of demand from their end-user would not justify investment in a dedicated deployment or PNI.

4.3 Private Network Interconnect (PNI)

PNI is the model in which a CDN and an ISP meet in a public datacenter. Both parties bear the cost of colocation and power and physical connection in the datacenter. Additionally, the ISP is responsible for the cost of backhauling traffic within its network. The interconnection is a Settlement-Free Interconnection (SFI) between the parties. It is also dedicated to the end-users of the ISP. CDN will not direct end-users from other entities on the Internet to the PNI. However, while the CDN to ISP interface is dedicated to the latter, the underlying hardware resources are not. As a result, ISPs whose PNIs are connected to the same node can find themselves competing for the hardware resources on the node. This can be especially unfavorable during high-demand events.

This model might seem more cost-effective than the IX model, since it does not include the IX port fee. However, there is an interconnection setup cost, which represents the cost of the cross-connect, as well as the router port reservation cost. Therefore, in case the demand which would be served in this model is not sufficient, it might be more cost-efficient to serve it through an IX interconnection.

4.4 Transit Interconnection

Transit interconnection (also known as IP transit) is an interconnection model in which the ISP offers the CDN full connectivity. In this model, the ISP agrees to provide CDN access to any location on the Internet. This can come at an expense to the ISP, as it might involve using IP transit services that they procure themselves or peering interconnections sensitive to imbalance. Therefore, in this model, the ISP charges the CDN for the use of the interconnection. This service is commonly charged for Megabits per second (Mbps) at 95th percentile usage of the link. In addition to the transit cost mentioned above, the deployment of a node in this model also comes at the expense of public data-center colocation. It is a result of the need to deploy the caching node at a location where the transit provider, with whom the CDN wants to connect, is present.

Tab. 2 summarizes the information presented in this section.

Tab. 2: Summary of parameters impacted by the inter-connectivity models

Interconnection model	*Cost Components*	*Connectivity scope*
Partnership	p_{hw}	V_{isp}
IX	p_{hw}, p_{colo}, p_{xcon}	V_{ix}
PNI	p_{hw}, p_{colo}, p_{xcon}, p_{port}	V_{isp}
Transit	p_{hw}, p_{colo}, p_{bw}	V

5. EXAMPLE MODEL ENHANCEMENT

In this section, we present an enhancement of the model presented in [11].

Eq. (1) brakes down p^l, known from the cited model, into more specific cases, depending on the interconnection model.

$$p^l = \begin{cases} p_{\text{hw}} & \text{if } l \in L_{\text{pship}} \\ p_{\text{hw}} + p^l_{\text{colo}} + (p^l_{\text{port}} + p^l_{\text{xcon}}) \left\lceil \frac{T_s x_k^{lV_{\text{ix}}}}{100} \right\rceil | x_k^{lV_{\text{ix}}} & \text{if } l \in L_{\text{ix}} \\ p_{\text{hw}} + p^l_{\text{colo}} + p^l_{\text{xcon}} \left\lceil \frac{T_s x_k^{lv}}{100} \right\rceil | x_k^{lv} & \text{if } l \in L_{\text{pni}} \\ p_{\text{hw}} + p^l_{\text{colo}} + 1000 p^l_{\text{bw}} T_s & \text{if } l \in L_{\text{transit}} \end{cases} \tag{1}$$

Sets denoted as variants of L with lower index group locations l, where different interconnection models are available. For example, L_{ix} groups the locations where it is possible to connect with ISPs via the peering infrastructure of an Internet Exchange point. In one physical facility, there may be different ways of interconnecting available. L_{transit} groups the locations where it is possible to connect with ISP in the Transit model. In the enhanced representation of the model, l should be understood as a combination of the physical location and the type of interconnection used to reach end-users.

The cost components have already been discussed in Section 4.

The p^l_{port}, price per IX port and p^l_{xcon}, price per cross-connect, are multiplied by the ceiling of x_k^{lv} and $x_k^{lV_{\text{ix}}}$, the number of servers serving end-users via the PNI and IX connection, respectively. This is then multiplied by each server's ability to deliver traffic and divided by 100. The 100 seen in Eq. 1 is an assumed port size - 100 Gbps. This is then divided by x_k^{lv} and $x_k^{lV_{\text{ix}}}$. This notation enables one to capture per-server cost of port and cross-connect. We evaluate the number of servers dedicated to serving end-users via a given interconnection model at a location l. Then determine how much traffic this would represent and translate it to the necessary number of ports and cross-connects.

In the Transit model, the per-server cost of deployment consists of hardware, colocation, and power cost in addition to the cost of the IP transit service. The latter is captured as a product of the cost per Mbps and the server's computational ability to deliver traffic expressed in Gbps and multiplied by 1000. Multiplication is performed to adjust the units to Mbps. In this model, we assume that the cost of cross-connect is factored in the cost of the IP transit service.

In Eqs. (2)–(14) we present the cited model enhanced with the insights oriented from different CDN-ISP connectivity models. For ease of discussing the significance of individual elements of the objective function and constraints, we break it into individual lines and elaborate on them separately.

$$\min \sum_{k=0}^{K} \sum_{v \in V_{\text{isp}}} \sum_{l \in L_{\text{pship}}} x_k^{lv} p_{k,hw} + \tag{2}$$

$$\sum_{v \in V_{\text{ix}}} \sum_{l \in L_{\text{ix}}} x_k^{lv} (p_{k,hw} + p^l_{k,colo} + \tag{3}$$

$$(p^l_{\text{xcon}} + p^l_{\text{port}}) \left\lceil \frac{T_s x_k^{lV_{\text{ix}}}}{100} \right\rceil | x_k^{lV_{\text{ix}}}) + \tag{4}$$

$$\sum_{v \in V_{\text{isp}}} \sum_{l \in L_{\text{pni}}} x_k^{lv} (p_{k,hw} + p^l_{k,colo} + \tag{5}$$

$$p^l_{\text{xcon}} \left\lceil \frac{T_s x_k^{lv}}{100} \right\rceil | x_k^{lv}) + \tag{6}$$

$$\sum_{v \in V} \sum_{l \in L_{\text{transit}}} x_k^{lv} (p_{\text{hw}} + p^l_{\text{colo}} + 1000 p^l_{\text{bw}} T_s) + \tag{7}$$

$$c^l (u_k^{lv})^2 \tag{8}$$

$$\text{s.t.} \sum_{l \in L} \frac{x_k^{lv}}{a^{lv}} \geq D_k^{lv}, \qquad \forall v \in V, 0 \leq k \leq K \tag{9}$$

$$x_{k+1}^{lv} = x_k^{lv} + u_k^{lv}, \qquad \forall l \in L, v \in V, 0 \leq k < K \tag{10}$$

$$\sum_{v \in V} x_k^{lv} \leq C^l, \qquad \forall l \in L, k \in K, \tag{11}$$

$$x_k^{lv} \in \mathbb{R}_{\geq 0} \setminus (0, x^l_{\min}), \forall l \in L, v \in V, 0 \leq k \leq K \tag{12}$$

$$u_k^{lv} \in \mathbb{R}, \qquad \forall l \in L, v \in V, 0 \leq k \leq K \tag{13}$$

$$E \subset \bigcup_{l \in L} \{l, V_l\} \tag{14}$$

Eqs. (2) - (8) consist of the objective function of

the enhanced problem formulation. The objective function has been enhanced by adopting the understanding of p^l presented in Eq. (1). Eq. (2) consists of the summarizing operation in all time steps. In addition, it denotes part of the objective function that verifies variants of deploying caching servers in the partnership model - it is done for all of the locations l where the partnership model is available and all the end users that can be connected in this model from the facility. In this case, the per-server cost of deployment is limited to the hardware cost.

Eqs. (3)-(4) denote the part of the objective function that verifies variants of deploying caching servers in the IX model. It is done for all locations l where the IX model is available and all end-users that can be connected in this model from the facility. In this case, the per-server cost of deployment consists of hardware and colocation. In addition to those, in this model, the CDN provider has to cover the costs which are dependent on the traffic level, i.e., port and cross-connect fees. These latter costs increase discretely with each new port added. In our formulation, the port size is assumed to be 100 Gigabits per second.

Eqs. (5)-(6) denote the part of the objective function that verifies variants of the deployment of caching servers in the PNI model. It is done for all locations l where the PNI model is available and all end users that can be connected in this model from the facility. In this case, the per-server cost of deployment consists of hardware, colocation, and cross-connect. The per-server cross-connect cost is captured in the same way as in case of the IX.

Eq. (7) denotes the part of the objective function that verifies variants of deploying caching servers in the transit model. It is done for all locations l where the Transit model is available and all end-users. In this case, the per-server cost of deployment consists of hardware, colocation, and transit cost charged per-Mbps.

Eqs. (8) - (11) and (13) do not vary with respect to the cited model. Their meaning described in [11] holds.

Eq. (13) has been modified in two ways: 1) extending the belonging of x_k^{lv} with 0; The cited model described the state variable as positive real numbers and assumed rounding up in case of fractions. As a result, every location would have at least one server allocated. We believe that there are cases where such an approach would result in unjustified expenditures. 2) Removal of the open interval from 0 to $x_{\min}^l$. The logic for such a modification is to factor in the target caching ratio of caching nodes resulting from the optimization. Depending on the content base and its popularity profile, the CDN operator might need more than one server worth of disk space to achieve the caching ratio they are targeting.

Eq. (14) summarizes the constraint that comes from the fact that not all end-user groups can be reached from all facilities. The edges of the bipartite graph have to belong to the set of edges connecting each of the vertices representing facilities and the set of vertices representing end-users, reachable from that facility.

6. CONCLUSIONS

In this paper, we presented practical models of interconnecting between ISPs and CDNs and how significantly different they are. We discussed the impact of each of them on the components that contribute to the cost of deployment of a CDN node in each of them and the scope connectivity available. We showed how a mature optimization model can be enhanced to capture specifics of the models. An important takeaway from the enhancement process is that while some of the deployment models might seem more attractive than others — e.g. having less cost components — they come with a minimal cost (e.g. minimal number of servers or cost of a cross-connect), which makes the model financially unattractive in case of low demand being served from using them. Therefore, deployment decisions must always be made with an understanding of available options and the specificity of the particular location. We believe that the research presented will support further refinement of existing models.

LITERATURA

[1] Akamai. 2023. "Network Operator | Akamai Network Partnerships.". URL https://www.akamai.com/why-akamai/network-operator/akamai-network-partnerships.

[2] Google. 2023. "Introduction to GCC - Interconnect Help". URL https://support.google.com/interconnect/answer/9058809?hl=en

[3] Huang Bo, Du Changlai, Sun Mengting, Yuan Xiaoqun. 2019. "A dynamic virtual datacenter selection strategy for integrated cloud service platform construction with multiclouds". IEEE Access 7 : 172879–172891.

[4] Jin Zizhe, Pan Li, Liu Shijun, 2021, "Online cost-effective edge service renting for content providers in cloud and edge environments". In 2021 IEEE International Conference on Web Services (ICWS) : 541–550.

[5] Moreno Muro Francisco-Javier, Skorin-Kapov Nina, Pavon-Marino Pablo. 2019. "Revisiting core traffic growth in the presence of expanding CDNs". Computer Networks, 154:1–11.

[6] Netflix,2023, "Netflix open connect". URL https://openconnect.netflix.com/en/.

[7] Nguyen Thanh Vinh, Chou Chun Tung, Boustead P. 2003. "Provisioning content distribution networks over shared infrastruc ture." In The 11th IEEE International Conference on Networks, ICON2003 : 119–124

[8] Nygren Erik, Sitaraman Ramesh K., Sun Jennifer. 2010. "The akamai network: A platform for high-performance internet applications". SIGOPS Oper. Syst. Rev., 44(3) : 2–19,

[9] Sahoo Jagruti, Salahuddin Mohammad A., Glitho Roch, Elbiaze Halima, Ajib Wessam. 2017. "A survey on replica server placement algorithms for content delivery networks". IEEE Communications Surveys Tutorials 19(2): 1002–1026.

[10] Xu Kai, Li Xiang, Bose Sanjay Kumar, Shen Gangxiang. 2018. Joint replica server placement, content caching, and request load assignment in content delivery networks. IEEE Access 6:17968–17981.

[11] Zhang Qi, Zhu Quanyan, Zhani Mohamed Faten, Boutaba Raouf. 2012. "Dynamic service placement in geographically distributed clouds". In 2012 IEEE 32nd International Conference on Distributed Computing Systems : 526–535.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

ZARZĄDZANIE WIELOWARSTWOWYMI ROZPROSZONYMI ZASOBAMI W SIECIACH 5G-ADVANCED I 6G Z ZACHOWANIEM BEZPIECZEŃSTWA I PRYWATNOŚCI DANYCH

ORCHESTRATION OF MULTI-LAYERED RESOURCES IN 5G-ADVANCED AND 6G NETWORKS IN A DISTRIBUTED ARCHITECTURE WITH DATA SECURITY AND PRIVACY

Tomasz Wichary[1]; Jordi Mongay Batalla[2]

[1,2] Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska,
{[1]tomasz.wichary.dokt; [2]jordi.mongay.batalla}@pw.edu.pl

DOI: 10.15199/59.2023.4.91

Streszczenie**: Następne generacje sieci mobilnych 5G-Advanced, 6G stwarzają nowe możliwości zastosowań dla przemysłu, gdzie wykorzystywane są inteligentne systemy sterowania aplikacjami przetwarzające informacje w czasie rzeczywistym. Przetwarzanie brzegowe umożliwia przybliżenie aplikacji do użytkownika końcowego np. urządzeń Internetu rzeczy i jednocześnie wprowadzania wysoce rozproszoną architekturę i złożoność operacyjną. Cały ekosystem wymaga orkiestracji od końca do końca, inteligentnego zarządzania i automatyzacji z uwzględnieniem aspektów bezpieczeństwa. W tym artykule przedstawiamy model reprezentacji zasobów w różnych warstwach oraz ich wzajemne relacje dla konkretnej realizacji sieci.**

Abstract**: New generations of 5G-Advanced, 6G mobile networks are creating new application opportunities for industry with intelligent application control systems that process information in real time. Edge processing enables applications to be brought closer to the end user, e.g., Internet of Things devices, while introducing highly distributed architecture and operational complexity. The entire ecosystem requires E2E orchestration, intelligent management, and automation with security constraints.**

Słowa kluczowe**: bezpieczeństwo i prywatność danych, przetwarzanie brzegowe, relacje między zasobami, zarządzanie sieciami mobilnymi, zasoby wielowarstwowe.**

Keywords**: E2E orchestration, edge computing, multilayer resources, resources relations, security, and privacy constraints.**

1. WSTĘP

Sieci 5G-Advanced/6G wprowadzają szereg możliwości dla aplikacji mobilnych dzięki oferowanej dużej przepustowości i niskim opóźnieniom. Zarządzenie takimi sieciami staje się coraz bardziej skomplikowane i wymaga uwzględnienia wielu czynników oraz skoordynowania wielu zasobów. Aplikacje często wymagają zaawansowanych rozwiązań z zakresu bezpieczeństwa i skalowalności. W miarę wzrostu liczby użytkowników lub danych, infrastruktura może wymagać odpowiedniego zarządzania zasobami, a uwierzytelnianie, autoryzacja, szyfrowanie danych i ich monitorowanie, może dodatkowo skomplikować ten proces. Aplikacje tworzą treści w oparciu o semantyczne odwzorowania i pojęcia abstrakcyjne, następnie informacje przekazywane są pomiędzy użytkownikami [1]. W celu uproszczenia zarządzania skomplikowanymi aplikacjami, ważne jest wykorzystanie odpowiednich narzędzi i metodologii.

Sieć 5G została opracowana przez organizację 3GPP. Architekturę sieci oraz podstawowe funkcje zostały opublikowane w wydaniu 15 specyfikacji technicznych. Kolejne wydania 16 i 17 były ulepszaniem standardu 5G i polegały na wprowadzeniu nowych funkcji, usprawnienie wydajności sieci i obsługi nowych zastosowań. Sieć 5G jest oparta na usługach (ang. SBA), oddzieleniu płaszczyzny sterowania i użytkownika (ang. CUPS) oraz przetwarzania brzegowego Multi-Access-Edge Computing (ETSI ISG MEC) [2]. Kolejne wydanie 18 specyfikacji, nazywane jako sieć 5G-Advanced wprowadza rozszerzenia m.in. dotyczące przetwarzania brzegowego, gdy użytkownik jest zalogowany do sieci innego operatora. UE musi dedykować sesję dla routingu ruchu lokalnego i inną sesję do własnej sieci dla pozostałych usług [3].

Architektura sieci mobilnej musi podążać za wymaganiami aplikacji i zarządzanie siecią będzie miało charakter zarządzania hierarchicznych domen technologicznych [4] oraz federacyjnego, czyli zdecentralizowanego przetwarzania aplikacji i danych [5]. Aplikacje komunikują się z własnymi komponentami lub innymi aplikacjami rozporoszonymi w różnych częściach sieci. Taki model architektury może skutkować tym, że np. problemy bezpieczeństwa jednego elementu zasobu infrastruktury mogą się przenieść na inne jej elementy oraz aplikacje, które współdzielą ten sam element infrastruktury.

Zarządzenie takimi usługami komplikuje fakt, że taka sieć będzie i) budowana w oparciu o specjalizowane rozwiązania pochodzące od różnych dostawców oraz ii) współdzielona przez wielu dzierżawców.

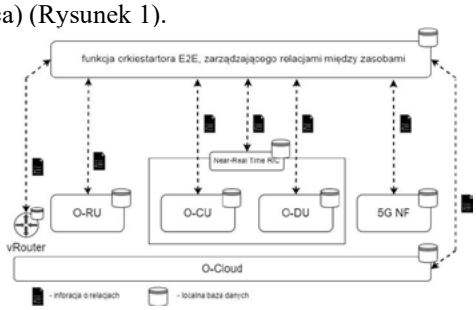
W sieciach mobilnych wyróżnia się domeny technologiczne, w szczególności sieć dostępową (ang. RAN), transportową i rdzeniową (ang. Core). Ze względu na lepsze wykorzystanie zasobów sprzętowych, domeny technologiczne mogą wykorzystywać zasoby zwirtualizowane i wprowadzają dodatkowe warstwy pomiędzy warstwą sprzętową i aplikacji. Przykładową siecią zbudowana na rozproszonej wielodomenowej infrastrukturze i rozwiązaniami od wielu dostawców jest sieć 5G-VINNI [6] . W sieci zostały zbudowane usługi bazujące na usłudze plastrów sieci (ang. network slicing [7]) opartym na

modelu Network Slice as a Service. Plastry sieci rozumiane są jako logiczne połączenie od urządzenia do serwerów świadczących usługę oraz sposób na współdzielenie zasobów infrastruktury.

Współdzielone zasobów sieciowych odbywa się w różnych domenach technologicznych i jest praktyką polegającą na udostępnianiu tych samych zasobów dla różnych użytkowników lub dzierżawców (ang. tenants). Udostępnianie zasobów wymaga *i*) zastosowania różnych mechanizmów bezpieczeństwa i polityki prywatności na całej ścieżce przesyłania danych jako kompleksowy sposób zabezpieczenia sieci 5G [8] oraz *ii*) odpowiedniego zarządzania zasobami w celu uniknięcia przeciążenia, utrzymywania odpowiedniej wydajności oraz poziomu bezpieczeństwa i prywatności danych dla wszystkich użytkowników. Dobór odpowiednich mechanizmów kontroli bezpieczeństwa może znacząco zredukować zagrożenie oraz wzmocnić wymagany poziom izolacji zasobów. W [9], autorzy proponują plastry technologicznie agnostyczne, które są tworzone w jednej domenie. Wielodomenowość zakłada wtedy wiele takich plastrów i możliwość tworzenia plastrów z innych plastrów.

System zarządzania zasobami powinien reagować na zmieniające się otoczenie i dostrajać parametry zasobów, protokołów w zależności od potrzeb aplikacji. Zakłada się, że tylko wybrane zasoby mogą być współdzielone oraz zasoby mogą być współdzielone przez wybrane usługi. Zależności między elementami w tej samej warstwie np. w warstwie sieciowej określone są przez procedury wymiany wiadomości sygnalizacyjnych, użytkownika oraz zarządzania i relacje pomiędzy nimi mogą być zdefiniowane przez dodatkowe warstwy.

Wszystkie informacje o zasobach we wszystkich warstwach będą dostępne w jednym ośrodku decyzyjnym. Tą funkcję będzie pełnić orkiestrator E2E relacji (tj. systemu, który zarządza zasobami w całej sieci od końca do końca) (Rysunek 1).



Rys. 1. Ośrodek decyzyjny (Orkiestrator E2E)

Orkiestrator będzie zbierał informacje dotyczące realizacji od zasobów sieciowych np. O-RU, O-DU, O-CU, 5GNF, vRouter, O-Cloud, RIC i podejmował decyzje związane z zasobami i relacjami np. w przypadku gdy wykryto podatność zasobu lub zasób jest atakowany.

W tym artykule przedstawiamy model powiązań zasobów w wielodomenowej sieci zbudowanej w oparciu o rozwiązania pochodzące od wielu dostawców i dzierżawców. Model ten przedstawiony jest w Sekcji 2. W sekcji 3, przykład zużycia modelu jest przedstawiony dla reorganizacji zasobów z podatnością dla jednej Funkcji Sieciowej (ang. Network Function, NF). Sekcja 4 podsumuje artykuł.

2. KONCEPCJA RELACYJNEJ ORKIESTRACJI PREDYKCYJNEJ

Do modelu powiązań pomiędzy zasobami wprowadzamy pojęcia: zasób oraz relację (pomiędzy zasobami).

2.1. Zasób

Jako zasób w sieci mobilnej przyjmujemy środki umożliwiające przetwarzanie danych w różnych płaszczyznach, np. w warstwie fizycznej, gdzie mamy urządzenia nadawczo-odbiorcze, serwery świadczącą usługę, w warstwie sieciowej, gdzie rozważamy funkcje sieciowe NF lub plastry sieci umożliwiających realizacje usługi od końca do końca oraz warstwy innych technologii takie jak zwirtualizowane zasoby definiowane jako pamięć masowa, moc obliczeniowa, łącze sieciowe, przekaźnik, bazy danych lub analitykę. Komunikacja pomiędzy urządzeniami składa się z etapów, które nazywane są warstwami. Warstwy grupują zasoby o podobnych właściwościach takich jak protokoły i mechanizmy komunikacji. Przyjmujemy, że każdy zasób w jednej warstwie formuje relację z zasobami w innych warstwach. Zasoby w różnych warstwach będące w relacji z innymi zasobami tworzą zasoby wielowarstwowe. Ogólny model sieci składa się z I warstw i każda z nich zawiera $J_i, i = 1..I$ zasobów. Każdy zasób definiujemy jako $O_{ij}, i = 1..I, j = 1..J_i$ (Rysunek 2). Jako przykład warstwy może być warstwa 1 reprezentująca plastry sieciowe. Każdy zasób tej warstwy to jeden plaster sieciowy. Kolejną warstwą może być warstwa funkcji sieciowych, gdzie każdy zasób tej warstwy byłby jedną instancją danej funkcji sieciowej. Trzecią warstwą mógłby być zbiór zwirtualizowanych węzłów klastra (model chmurowy). Warstwą czwartą mogą być np. PODy, czyli najmniejsze zarządzane obiekty, które można tworzyć i zarządzać w platformie Kubernetes.

2.2. Relacja

Mówimy, iż zasób j w warstwie i $(O_{i,j})$ ma relację z zasobem k w warstwie m $(O_{m,k})$ kiedy działanie zasobu $O_{i,j}$ wymaga działania zasobu $O_{m,k}$. Proszę brać pod uwagę iż model ten reprezentuje stan zasobów sieci w danej chwili. Zasoby i relacje mogą się zmieniać w czasie.

Przyjmiemy następującą notację do opisu relacji pomiędzy zasobami w różnych warstwach.
$P_{i,j,k}$ – oznaczenie relacji pojedynczego elementu j w warstwie i do elementu k w kolejnej warstwie $i+1$:

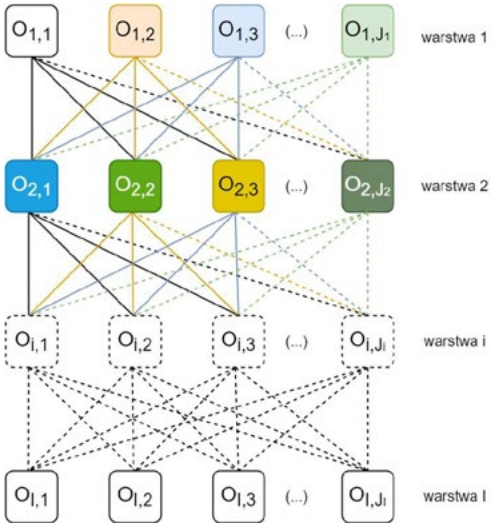
$$P_{i,j,k} \in \{0,1\}, 1 \le i < I, 1 \le j \le J_i, 1 \le k \le J_{i+1} \quad (3)$$

oraz $P_{i,j,k} = \begin{cases} 0 \leftrightarrow \textit{nie ma relacji między zasobem j w warstwie i oraz zasobem k w warstwie } i+1 \\ 1 \leftrightarrow \textit{jest relacja między zasobem j w warstwie i oraz zasobem k w warstwie } i+1 \end{cases}$

$P_{i,j}$ – to wektor (J_{i+1} elementów $\in \{0,1\}$) pokazujący istnienie relacji między zasobem j w warstwie i oraz wszystkimi zasobami w warstwie $i+1$.

$$P_{ij}^T = [P_{i,j,1}, P_{i,j,2}, \dots, P_{i,j,J_{i+1}}] \quad (4)$$

Wtedy macierz relacji między zasobami w warstwie i oraz warstwie $i+1$ będzie następująca:
$M_{i,i+r}$– postać macierzowa relacji pomiędzy wszystkimi zasobami w warstwie i oraz wszystkimi zasobami $i+r$. Proszę zauważyć, że macierz $M_{i,i+r}$ będzie miała J_i rzędów oraz J_{i+r} kolumn, tj. $M_{i,i+r} = (a_{rs})_{1 \le r \le J_i, 1 \le s \le J_{i+r}}$.



Rys. 2. Reprezentacja wielowarstwowych zasobów

Do przeliczenia macierzy $M_{i,i+r}$ orkiestrator E2E potrzebuje informacji o relacjach między zasobami w warstwie i oraz zasobami warstwie i+1. Orkiestrator E2E otrzymuje informację od orkiestratora niższego rzędu na poziomie funkcji sieciowych (standaryzowane przez 3GPP TS 28.533[10]) oraz od orkiestratorów odpowiadających za zarządzaniem maszynami wirtualnymi oraz węzłami i PODami (standaryzowane przez ETSI dokument NFV-MAN 001 [11-12]). Orkiestrator E2E współpracuje z platformami Openstack i Kubernetes i otrzymuje informację o tym gdzie dana funkcja sieciowa jest instalowana oraz w jakich węzłach pracują PODy.

Dlatego definiujemy wektor O_i jako zbiór wszystkich zasobów warstwy i.

$$O_i = \begin{bmatrix} O_{i,1} \\ O_{i,2} \\ \vdots \\ O_{i,J_i} \end{bmatrix}, \quad (5)$$

Rozważmy trzy zbiory zasobów O_1, O_2 i O_3 odpowiednio w warstwach 1, 2 i 3. Relacje pomiędzy warstwami 1 i 2 oznaczymy jako złożenie relacji $O_2 o O_1$, relacje pomiędzy warstwami 2 i 3 jako złożenie relacji $O_3 o O_2$ oraz relacje pomiędzy warstwami 1 i 3 jako złożenie relacji $O_3 o O_1$.

Relacje pomiędzy warstwami 1 i 2 ($O_2 o O_1$) są relacjami dwuargumentowymi, które można zdefiniować:

$$O_2 o O_1 = \bigcup_{r=1}^{J_1} \bigcup_{s=1}^{J_2} P_{1,r,s} \quad (6)$$

Dziedziną tej relacji będzie zbiór

$$D_1(O_2 o O_1) = \left\{ O_{1j_1} \epsilon O_1 : \exists_{O_{2j_2} \epsilon O_2} \, (O_{1j_1}, O_{2j_2}) \epsilon O_2 o O_1 \right\}, \quad (7)$$

natomiast obrazem tej relacji jest zbiór

$$D_2(O_2 o O_1) = \left\{ O_{2j_2} \epsilon O_2 : \exists_{O_{1j_1} \epsilon O_1} \, (O_{1j_1}, O_{2j_2}) \epsilon O_2 o O_1 \right\}, \quad (8)$$

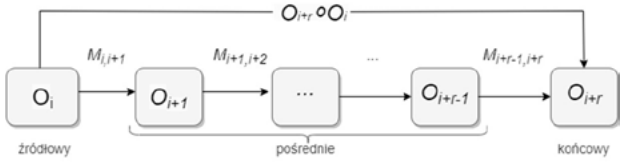
Relacje pomiędzy warstwami 2 i 3 są relacjami dwuargumentowymi $O_3 o O_2$. Dziedziną tej relacji będzie zbiór

$$D_2(O_3 o O_2) = \left\{ O_{2j_2} \epsilon O_2 : \exists_{O_{3j_3} \epsilon O_3} \, (O_{2j_2}, O_{3j_3}) \epsilon O_3 o O_2 \right\}, \quad (9)$$

natomiast obrazem tej relacji jest zbiór

$$D_3(O_3 o O_2) = \left\{ O_{3j_3} \epsilon O_3 : \exists_{O_{2j_2} \epsilon O_2} \, (O_{2j_2}, O_{3j_3}) \epsilon O_3 o O_2 \right\}, \quad (10)$$

Rozważmy następujące zbiory: zbiór źródłowy O_i; zbiory pośrednie O_n, gdzie $i < n < i + r$; oraz zbiór końcowy O_{i+r}.



Rys. 3. Zespolenie relacji między warstwami.

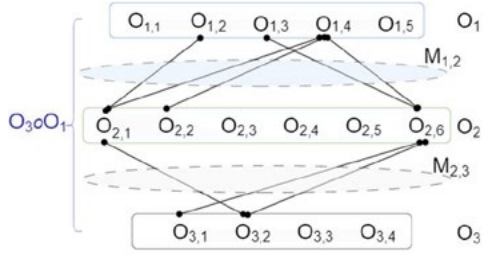
Relacja $O_{i+r} o O_i$ jest zespoleniem relacji dwuargumentowych, a zespoleniem relacji $O_{i+r} o O_i$, będzie produkt macierzowy (Rysunek 3):

$$M_{i,i+r} = M_{i,i+1} x M_{i+1,i+2} x \ldots x M_{i+r-1,i+r}. \quad (11)$$

3. ZASOBY Z PODATNOŚCIĄ

W przypadku tego modelu, każdy użytkownik otrzymuje określoną część zasobów. Oceną podatności zajmuje się Orchestrator E2E, analizuje i ocenia potencjalny wpływ jaki może mieć podatność. Wpływ może być zależny od poziomu dostępu, typu danych np. danych wrażliwych, zakresu i jakie zagrożenie niesie podatność.

Na rysunku 3 przedstawiamy sieć składającą się z pewnych usług w postaci plastrów sieci. Plastry sieci wymagają zasobów funkcji sieciowych oraz zasobów węzłów platformy Kubernetes do realizacji swoich usług.



Rys. 4. Przykład relacji zasobów w różnych warstwach.

W przykładzie zakładamy sieć zbudowaną z zasobów wielowarstwowych, która składa się z następujących warstw: w warstwie 1 znajduje się 5 zasobów $\{O_{1,1}, O_{1,2}, O_{1,3}, O_{1,4}, O_{1,5}\}$ reprezentujących 5 plastrów sieciowych. W warstwie 2 znajduje się 6 zasobów $\{O_{2,1}, O_{2,2}, O_{2,3}, O_{2,4}, O_{2,5}, O_{2,6}\}$ reprezentujących 6 funkcji sieciowych. W tym przykładzie funkcję sieciową cechuje uniwersalność, oznaczającą że każda funkcja sieciowa może zastąpić każda inną funkcję. Ostatnia warstwa zawiera 4 zasoby $\{O_{3,1}, O_{3,2}, O_{3,3}, O_{3,4}\}$ reprezentujące 4 węzły zarządzane przez klaster Kubernetes.

Orkiestrator E2E jest centralnym punktem, który zbiera informację o relacjach pomiędzy kolejnymi warstwami. Na tej podstawie oblicza relacje pomiędzy warstwami, które ze sobą nie sąsiadują, w przykładzie między warstwą 1 oraz warstwą 3.

Zdefiniujemy relacje dla każdego pojedynczego zasobu w warstwie 1 do wszystkich zasobów w warstwie 2 oraz relacje dla każdego pojedynczego zasobu w warstwie 2 do wszystkich zasobów w warstwie 3. Będzie to zbiór relacji $(O_{1,2}, O_{2,1})$, $(O_{1,3}, O_{2,6})$, $(O_{1,4}, O_{2,1})$, $(O_{1,4}, O_{2,2})$ oraz $(O_{1,4}, O_{2,6})$ pomiędzy wszystkimi zasobami warstw 1 i 2. Wektory relacji dla każdego zasobu w warstwie 1 do wszystkich zasobów w warstwie 2 będą: $P_{11}^T = [0\,0\,0\,0\,0\,0]$, $P_{12}^T = [1\,0\,0\,0\,0\,0]$, $P_{13}^T = [0\,0\,0\,0\,0\,1]$, $P_{14}^T = [1\,1\,0\,0\,0\,1]$, $P_{15}^T = [0\,0\,0\,0\,0\,0]$.

Te relacje tworzą złożenie realizacji, której postać macierzowa wygląda następująco

$$M_{1,2} = \begin{bmatrix} P_{11}^T \\ P_{12}^T \\ P_{13}^T \\ P_{14}^T \\ P_{15}^T \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Zbiorem relacji pomiędzy warstwami 2 i 3 będzie $(O_{2,1}, O_{3,2})$, $(O_{2,6}, O_{3,1})$ oraz $(O_{2,6}, O_{3,2})$.

Wektory relacji dla każdego zasobu w warstwie 2 do wszystkich zasobów w warstwie 3 będą wyglądały następująco: $P_{21}^T = [0\,1\,0\,0]$, $P_{22}^T = [0\,0\,0\,1]$, $P_{23}^T = [0\,0\,0\,0]$, $P_{24}^T = [0\,0\,0\,0]$, $P_{25}^T = [0\,0\,0\,0]$, $P_{26}^T = [1\,1\,0\,0]$.

Te relacje tworzą złożenie realizacji, której postać macierzowa wygląda następująco

$$M_{2,3} = \begin{bmatrix} P_{21}^T \\ P_{22}^T \\ P_{23}^T \\ P_{24}^T \\ P_{25}^T \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}.$$

Złożeniem relacji $O_3 o O_1$, będzie postać macierzowa $M_{O_3 o O_1}$, gdzie

$$M_{O_3 o O_1} = M_{1,2} x M_{2,3} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

oraz złożenie relacji w postaci zbioru

$$O_3 o O_1 = \left\{ \begin{matrix} (O_{1,2}, O_{3,2}), (O_{1,3}, O_{3,1}), (O_{1,3}, O_{3,2}), \\ (O_{1,4}, O_{3,1}), (O_{1,4}, O_{3,2}) \end{matrix} \right\}.$$

Relacja $(O_{1,4}, O_{2,2})$ występująca na rysunku 3 jest relacją szczególną, ponieważ z zasobu $O_{2,2}$ nie mamy relacji do zasobów w kolejnej warstwie 3. Dzieje się tak, ponieważ Orkiestrator E2E nie otrzymał jeszcze informacji o przydziale zasobów w niższych warstwach i możliwych relacjach pomiędzy warstwami.

Zauważmy, że w przypadku wykrycia zasobu z podatnością w warstwie drugiej, np. $O_{2,1}$, wszystkie relacje $(O_{1,2}, O_{2,1})$, $(O_{1,4}, O_{2,1})$, $(O_{2,1}, O_{3,2})$ muszą zostać usunięte. W wyniku rekonfiguracji sieci, te relacje zostaną zastąpione nowymi relacjami $(O_{1,2}, O_{2,x'})$, $(O_{1,4}, O_{2,x''})$, $(O_{2,x^*}, O_{3,2})$, gdzie x^* może wskazywać zasób $O_{2,x'}$ lub $O_{2,x''}$ w warstwie 2. Rekonfiguracja sieci nie ma wpływu relacje między warstwami 1 i 3 i $M_{O_3 o O_1}$ nie zmieni się.

Zauważmy również, że wykrycie podatności w warstwach pierwszej i ostatniej (np. w warstwie 1 i 3) powoduje konieczność przebudowania relacji tylko z warstwą najbliższą (warstwa 2). Skomplikowanie tej operacji zależy od liczby istniejących relacji.

4. PODSUMOWANIE

W artykule przedstawiliśmy koncepcję modelu powiązań pomiędzy zasobami w różnych warstwach sieci. Skupiliśmy się na opisie relacji pomiędzy bezpośrednimi warstwami oraz zasobami w warstwach niesąsiadujących ze sobą. Dla systemu zarządzania wiedza o współzależnościach zasobów w różnych warstwach pozwala na podjęcie odpowiednich kroków w ograniczaniu skutków ataków lub podatności zasobu. W szczególności różne działania muszą zostać podjęte w zależności gdzie została wykryta podatność i od ilości relacji z innymi zasobami.

PODZIĘKOWANIE

Ten artykuł był dofinansowany przez Narodowe Centrum Badań i Rozwoju w ramach projektu Nr. CYBERSECIDENT/489818/IV/NCBR/2021.

LITERATURA

[1] Park, Sang-Min, and Young-Gab Kim. 2022. “A Metaverse: Taxonomy, Components, Applications, and Open Challenges.” IEEE Access 10: 4209–51. https://doi.org/10.1109/ACCESS.2021.3140175.

[2] “Multi-access Edge Computing (MEC); Support for network slicing”, 2019, ETSI GR MEC 024 V2.1.1, https://www.etsi.org/deliver/etsi_gr/MEC/001_099/024/02.01.01_60/gr_mec024v020101p.pdf

[3] “Becoming 5G Advanced the 3GPP 2025 Roadmap”, 2022, https://www.5gamericas.org/wp-content/uploads/2022/12/Becoming-5G-Advanced-the-3GPP-2025-Roadmap-InDesign.pdf.

[4] Dryjanski, Marcin, and Adrian Kliks. “A Hierarchical and Modular Radio Resource Management Architecture for 5G and Beyond.” IEEE Communications Magazine 58, no. 7 (July 2020): 28–34. https://doi.org/10.1109/MCOM.001.1900796.

[5] Coutinho, Rodolfo, Azzedine Boukerche. 2022 “Design of Edge Computing for 5G-Enabled Tactile Internet-Based Industrial Applications.” IEEE ComMag 60(1)60–66, https://doi.org/10.1109/MCOM.001.21261.

[6] Ordonez-Lucena, Jose, Christos Tranoris, and João Rodrigues. 2020. “Modeling Network Slice as a Service in a Multi-Vendor 5G Experimentation Ecosystem.” In 2020 IEEE International Conference on Communications Workshops (ICC Workshops), 1–6. https://doi.org/10.1109/ICCWorkshops49005.2020.9145225.

[7] “Description of Network Slicing Concept.,” 2016. https://ngmn.org/wp-content/uploads/160113_NGMN_Network_Slicing_v1_0.pdf.

[8] Wichary, Tomasz, Jordi Mongay Batalla, Constandinos X. Mavromoustakis, Jerzy Żurek, and George Mastorakis. 2022. “Network Slicing Security Controls and Assurance for Verticals.” Electronics 11, no. 2: 222. https://doi.org/10.3390/electronics11020222.

[9] Kukliński, Slawomir, Lechosław Tomaszewski, Robert Kołakowski, and Prosper Chemouil. 2021. “6G-LEGO: A Framework for 6G Network Slices.” Journal of Communications and Networks 23, no. 6: 442–53. https://doi.org/10.23919/JCN.2021.000025.

[10] Technical Specification Group Services and System Aspects; Management and orchestration; Architecture framework, document TS28.533 V17.3.0, 3GPP, 2023

[11] Network Functions Virtualisation (NFV); Management and Orchestration, document GS NFV-MAN 001 V1.2.1, ETSI, 2021.

[12] Choi, Jin Seek, Se Joon Chun, and Sunghwan Lee. “Hierarchical Distributed Overarching Architecture of Decoupled Federation and Orchestration Frameworks for Multidomain NFV MANOs.” IEEE ComMag 60 (9): 68–74. 2022. https://doi.org/10.1109/ MCOM.005.20949.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

Redukcja wielkości tablicy przepływów dzięki detekcji dużych przepływów za pomocą uczenia maszynowego

Flow table occupancy reduction based on machine learning elephant flow detection

Bartosz Kądziołka[1]; Piotr Jurkiewicz[2];

[1] AGH Akademia Górniczo-Hutnicza, Kraków, kadziolka@agh.edu.pl
[2] AGH Akademia Górniczo-Hutnicza, Kraków, piotr.jurkiewicz@agh.edu.pl

DOI: 10.15199/59.2023.4.92

Streszczenie: Liczba jednoczesnych przepływów w sieciach nadal przekracza pojemność tablic przepływów. W celu zmniejszenia zajętości tablic przeanalizowano wybrane modele uczenia maszynowego, wytrenowane na replikowalnych, rzeczywistych modelach ruchu, aby klasyfikować przepływy już od pierwszego pakietu. Jak pokazano w tej pracy, możliwe jest zmniejszenie liczby wpisów w tablicach 30-50 krotnie, przy jednoczesnym zachowaniu 80% pokrycia ruchu.

Abstract: The number of simultaneous flows in networks still overwhelms the capacities of the flow tables. Selected machine learning models trained on the reproducible, real traffic models to classify flows since the first packet were analyzed in order to reduce the flow table occupancy. As it is shown in this paper it is possible to reduce the number of flow entries by a factor up to 30-50, still covering 80% of the traffic using out-of-the-box models.

Słowa kluczowe: przepływy, tablica przepływów, optymalizacja, inżynieria ruchu, uczenie maszynowe

Keywords: flows, flow table, optimization, traffic engineering, machine learning

1. WSTĘP

Przyjmuję się, że rozkład długości oraz wielkości przepływów w Internecie podlega zasadzie Pareto, tzn. że niewielka liczba przepływów odpowiada za znaczącą większość ruchu. Ostatnie analizy wskazują, że za 80% całkowitego ruchu odpowiada jedynie 0,2-0,4% przepływów [10][12].

W trasowaniu przepływowym (*flow-based routing*) każdy przepływ posiada swój indywidualny wpis w pamięci przełącznika zawierający kolejny skok. Dzięki temu różne przepływy pomiędzy tymi samymi węzłami mogą podążać którąkolwiek z dostępnych tras. Ponadto, trasy dla kolejnych przepływów mogą być wybierane w zależności od aktualnego rozkładu obciążenia w sieci. Pozwala to na osiągnięcie transmisji adaptacyjnej i wielościeżkowej.

Niestety, liczba jednoczesnych przepływów w sieci jest znacznie większa niż pojemność tablic przepływów w przełącznikach [16]. Dodatkowym problemem w przypadku sieci scentralizowanej SDN jest ograniczona przepustowość kontrolera w obsłudze nowych przepływów. Rozwiązaniem tych kwestii mogłoby być tworzenie indywidualnych wpisów tylko i wyłącznie dla największych przepływów (*elephant flows*), podczas gdy pozostała większość małych przepływów (*mice flows*) była by trasowana po najkrótszych ścieżkach. Pozwoliłoby to na znaczne zredukowanie liczby wpisów w tablicach, zapewniając przy tym pokrycie znaczącej części ruchu kompleksowymi metodami inżynierii ruchu (TE).

Problemem pozostaje jak wykrywać największe przepływy. Korzystne jest wykrywanie ich jak najwcześniej, aby szybko utworzyć dla nich indywidualne wpisy i jak największą część ich pakietów objąć indywidualną inżynierią ruchu. Najbardziej korzystne byłoby klasyfikowanie przepływu wraz z jego pierwszym pakietem. Pozwoliłoby to także na uniknięcie przetrasowywania (*rerouting*) przepływu w trakcie jego trwania.

W tej pracy przeanalizowane zostały metody uczenia maszynowego (ML), w szczególności klasyfikatory oraz regresory w celu identyfikacji dużych przepływów. Za punkt odniesienia wybrana została metoda *first* [9]. Badania skupiają się na momencie identyfikacji przepływu, wielkości ruchu przesłanego w ramach zidentyfikowanego przepływu po jego identyfikacji oraz wynikającej z identyfikacji redukcji wielkości tablicy przepływów.

2. POWIĄZANE PRACE

Idea adaptacyjnego trasowania tylko największych przepływów została po raz pierwszy zaproponowana w 1999 roku [15]. Szczególnie w ostatnim czasie zyskała ona na popularności dzięki technologii SDN oraz uczenia maszynowego, gdzie kontroler bazując na wiedzy o całej sieci może odpowiednio zarządzać dużymi przepływami.

Hedera [1] to system w którym przepływy są przetrasowywane przez kontroler po przekroczeniu określonej wielkości. Urządzenia brzegowe zbierają statystyki wszystkich przepływów za pomocą liczników OpenFlow, zatem optymalizacja tablic dotyczy tylko pozostałych urządzeń. DevoFlow [2] to inny złożony system TE, który także skupia się na dużych przepływach - w celu ich detekcji używa próbkowania oraz liczników. Badanym parametrem wydajnościowym jest zagregowana przepływność sieci. W [18] duże przepływy są identyfikowane w urządzeniach brzegowych. W tej pracy przyjęto jednak model zakładający, że 20% przepływów odpowiada za 80% ruchu, co znacznie odbiega od rzeczywistości [10][12].

W powyższych pracach do detekcji dużych przepływów stosowane są proste metody takie jak próbkowanie, liczniki czy wartości progowe. Metody ML również są stosowane w tym celu. W [17] autorzy wykorzystują drzewo decyzyjne w celu identyfikacji dużych przepływów. Anali-

za skupia się głównie na dokładności (*accuracy*) identyfikacji. Praca [14] analizuje wydajność 3 różnych metod ML do predykcji wielkości przepływu. Analizowanymi parametrami były: procent prawidłowo rozpoznanych dużych przepływów (*true positive rate*) oraz procent prawidłowo rozpoznanych małych przepływów (*true negative rate*). W [11] oraz [6] zaproponowana została dwupoziomowa architektura klasyfikacji. Pre-klasyfikacja umiejscowiona jest w urządzeniach brzegowych, natomiast klasyfikacja w centralnym kontrolerze. W obydwu rozwiązaniach przebadana została w głównej mierze dokładność klasyfikacji.

We wszystkich wspomnianych pracach klasyfikacja przepływów odbywa się dopiero po kilku-kilkunastu pierwszych pakietach. Naszym celem jest tymczasem jak najszybsza klasyfikacja przepływu - już w jego pierwszym pakiecie. W [4] klasyfikacja bazuje wyłącznie na wielkości oraz nagłówku pierwszego pakietu - 5-tuple (*src IP, dst IP, src port, dst port, protocol*). W [3] autorzy zaproponowali predyktor wielkości oraz długości nowego przepływu wykorzystując *locally weighted regression*. Algorytm wykorzystuje zachowanie poprzednich przepływów oraz ich chwilową korelację z nowym przepływem. 5-tuple pierwszego pakietu jest również wykorzystywane w [8][7] przez głęboką sieć neuronową w celu predykcji charakterystyki przepływu. [5] jest najnowszą pracą skupiającą się na klasyfikacji już przy pierwszym pakiecie przepływu - autorzy analizują dokładność klasyfikacji.

Żadna ze wspomnianych prac nie analizuje jednak metryk takich jak redukcja tablicy przepływów czy ilość ruchu przesłanego po sklasyfikowaniu przepływu, które według nas są najważniejsze z perspektywy inżynierii ruchu i szeroko pojętego QoS. Skupiają się one na dokładności klasyfikacji, wyrażonej przez parametry takie jak *true positive rate*, *true negative rate* czy dokładność predykcji długości oraz wielkości przepływu. Niewiele mówią one jednak o skuteczności analizowanych algorytmów w określonym przez nas zastosowaniu. Przykładowo, błędna klasyfikacja największego przepływu w sieci ma znacznie większy wpływ na zmianę stopnia pokrycia ruchu niż błędne sklasyfikowanie przepływu małego. Dotychczas prezentowane metryki nie uwzględniają tej różnicy. Z kolei w artykułach proponujących kompleksowe systemy TE analizowane są metryki globalne, takie jak ilość przesłanego ruchu czy straty w sieci. Problem detekcji dużych przepływów na potrzeby redukcji wielkości tablic w systemach TE pozostaje zatem niezbadany w izolacji.

3. ALGORYTMY

Identyfikacja dużych przepływów ma znaczący wpływ na wielkość tablicy. Główne założenie tej pracy brzmi następująco: wpis w tablicy przepływów tworzony jest w momencie nadejścia pierwszego pakietu przepływu tylko i wyłącznie wtedy, gdy zostanie on sklasyfikowany jako przepływ duży. Klasyfikacja przepływu już przy pierwszym pakiecie jest możliwa i w tej pracy bazuje ona na dwóch typach uczenia maszynowego: klasyfikacji oraz regresji. Zarówno klasyfikacja jak i regresja to dwie główne metody uczenia nadzorowanego. Uczenie nadzorowane charakteryzuje się tym, że dane wejściowe muszą być odpowiednio oznaczone, aby model był w stanie dokonać klasyfikacji lub predykcji. Wszystkie modele których ewaluacja została przeprowadzona w tej pracy są dostępne w bibliotece scikit-learn[1].

[1]https://scikit-learn.org/

3.1 Klasyfikatory

W przypadku klasyfikatorów wyjściem algorytmu jest decyzja binarna czy przepływ jest duży czy nie. W czasie trenowania ze zbioru treningowego wybrany został procent największych przepływów, który daje określone pokrycie całego ruchu w sieci i dla nich trening odbywa się z decyzją 1, a dla pozostałych z decyzją 0. Następnie dla zbioru testowego klasyfikator przewiduje decyzję. Aby otrzymać krzywą redukcji wielkości tablicy w funkcji pokrycia, całą procedurę trenowania i dopasowywania trzeba powtórzyć dla różnych wartości oczekiwanego pokrycia ruchu (przyjęte zostało 25 różnych wartości).

3.2 Regresory

W przypadku regresorów wyjściem algorytmu jest przewidywana wielkość przepływu w bajtach. Aby otrzymać krzywą na wykresie nie trzeba powtarzać procedury treningu i dopasowania. Wystarczy zasymulować decyzję przyjmując różną wartość progową na przewidywanych wielkościach przepływu.

4. PRZYGOTOWANIE DANYCH

Zbiór danych ma kluczowy wpływ na wydajność badań opartych o metody ML. W tej pracy wykorzystany został zbiór danych agh_2015. Zbiór ten zawiera ponad 4 miliardy przepływów [10] zbieranych przez 30 dni.

Jako dane wejściowe wykorzystywane są informacje zawarte w nagłówku pierwszego pakietu. Sprawdzone zostały dwie różne reprezentacje danych wejściowych:

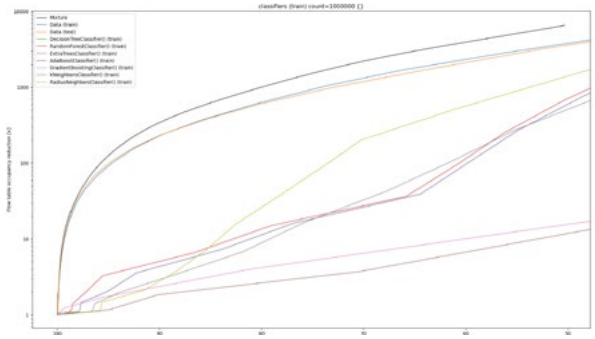
- pola nagłówka (*src IP, dst IP, src port, protocol*) w postaci takiej jak występują, co daje 4 cechy;
- te same pola nagłówka, przy czym pola większe niż 1 bajt zostały podzielone są na oktety, dając w sumie 11 cech, każda wielkości 8 bitów.

Dane zostały podzielone na zbiór treningowy oraz testowy obejmujący kolejno 80% oraz 20%. W trakcie analizy wielkości zbioru danych optymalny okazał się być zbiór zawierający 1 milion przepływów. Do obróbki danych został użyty pakiet [13]. Każdy algorytm został wytrenowany na danych treningowych, a następnie sprawdzony zarówno na tych samych danych treningowych jak i na danych testowych w celu weryfikacji wydajności.

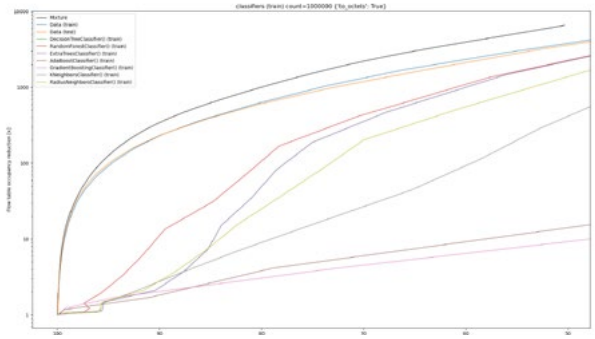
5. WYNIKI

Na wykresach poniżej zaprezentowano redukcję wielkości tablicy przepływów w funkcji pokrycia ruchu (oś x jest odwrócona, natomiast oś y logarytmiczna). Celem jest jak największa redukcja liczby wpisów, przy utrzymaniu jednocześnie jak największego pokrycia ruchu. Czarna linia (**mixture**) to obliczona wydajność dla rozkładu dopasowanego do pełnego zbioru danych (4 miliardy przepływów) przy założeniu że każdy przepływ został poprawnie sklasyfikowany. W tym celu wybrana została najmniejsza możliwa liczba przepływów w kolejności od największego, tak aby sumarycznie pokryła określony procent ruchu w sieci. Linia niebieska (**data (train)**) i linia pomarańczowa (**data (test)**) zostały obliczone w identyczny sposób do **mixture**, ale dla zbiorów treningowych i testowych używanych w konkretnej symulacji. Mogą być one zarówno poniżej jak i powyżej krzywej teoretycznej **mixture** (jeśli

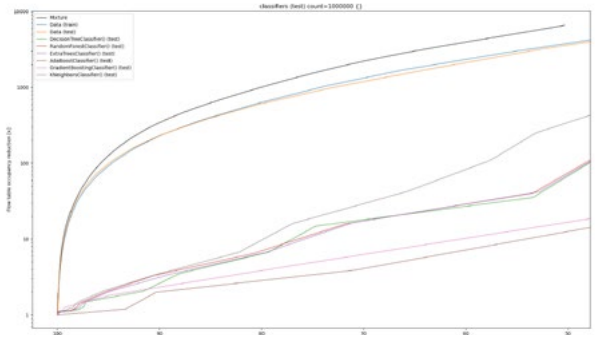
np. w zbiorze trafiła się nadreprezentacja dużych przepływów w stosunku do pełnego zbioru danych). Wyniki działań algorytmów należy zatem odnosić do tych krzywych niż do krzywej **mixture**.



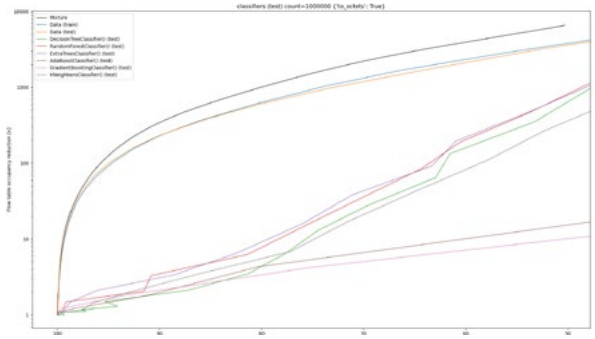
Rysunek 1: Klasyfikatory trening 4 cechy



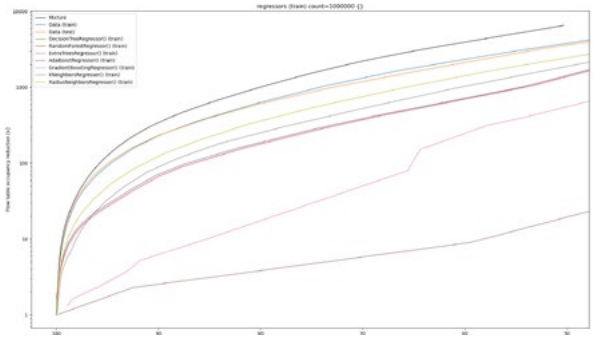
Rysunek 2: Klasyfikatory trening 11 cech



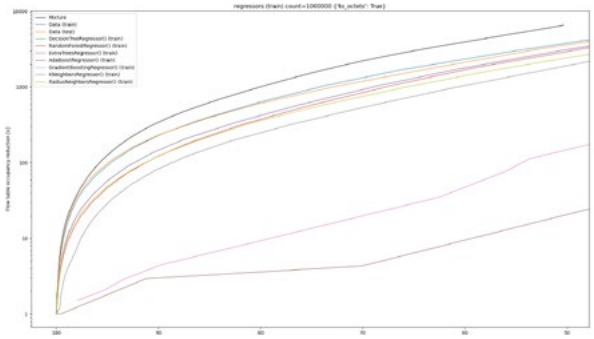
Rysunek 3: Klasyfikatory 4 cechy



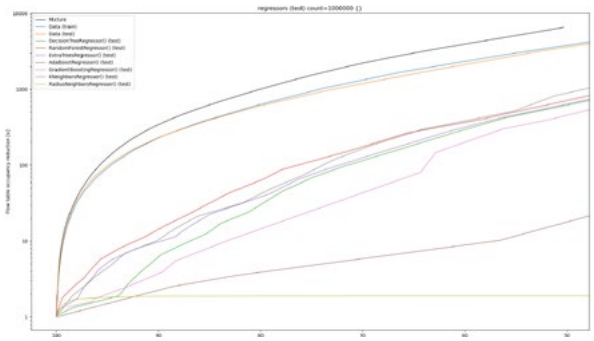
Rysunek 4: Klasyfikatory 11 cech



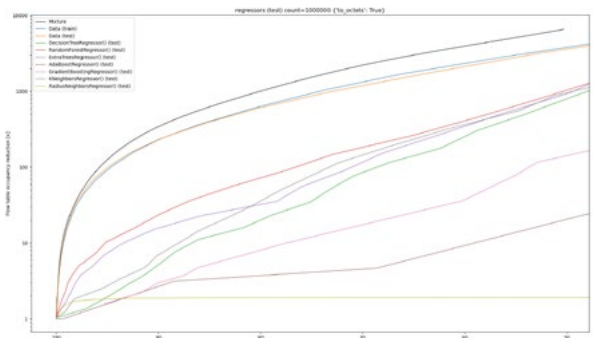
Rysunek 5: Regresory trening 4 cechy



Rysunek 6: Regresory trening 11 cech



Rysunek 7: Regresory 4 cechy



Rysunek 8: Regresory 11 cech

Widoczne na rysunkach 1, 2, 3 oraz 4 klasyfikatory nie radzą sobie dobrze. Jak widać są mało wydajne. Poza aspektem wydajnościowym sam trening jest problematyczny i zajmuje dużo czasu ponieważ model należy osob-

no wytrenować dla różnego pokrycia. Dodatkowo, klasyfikatory nie są wydajne nawet gdy poda im się ten sam zbiór danych na którym były trenowane, co przedstawiono na rysunkach 1 oraz 2. Spośród badanych modeli najlepiej prezentują się *RandomForest* oraz *ExtraTree* dla 11 cech oraz *KNeighbors* dla 4 cech. W tym przypadku liczba cech nie miała znaczącego wpływu na wyniki.

Regresory radzą sobie znacznie lepiej co widać na rysunkach 5, 6, 7 oraz 8. Na danych testowych najlepszą wydajność wykazuje model *RandomForest*, pomimo że na danych treningowych wypada najsłabiej, co oznacza iż pozostałe modele się przetrenowują. Przetworzenie danych wejściowych do oktetów ma pozytywny wpływ na wydajność szczególnie w obszarze pokrycia 100-80%.

Model *RandomForest* wydaje się najbardziej obiecującym kierunkiem badań. Przy pokryciu ruchu 80% pozwala na 30-50-krotną redukcję zajętości tablicy dla danych testowych.

6. DALSZE BADANIA

W przyszłości planowana jest bardziej szczegółowa analiza wybranych modeli ML, w tym sieci neuronowych. Modele zostaną wytrenowane na znacząco większym zbiorze danych. W przypadku sieci neuronowych planowane jest użycie biblioteki PyTorch[2] oraz procesorów graficznych.

LITERATURA

[1] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, A. Vahdat *et al.*, "Hedera: dynamic flow scheduling for data center networks." in *Nsdi*, vol. 10, no. 8. San Jose, USA, 2010, pp. 89–92.

[2] A. R. Curtis, J. C. Mogul, J. Tourrilhes, P. Yalagandula, P. Sharma, and S. Banerjee, "Devoflow: Scaling flow management for high-performance networks," in *Proceedings of the ACM SIGCOMM 2011 Conference*, 2011, pp. 254–265.

[3] M. V. B. da Silva, A. S. Jacobs, R. J. Pfitscher, and L. Z. Granville, "Predicting elephant flows in internet exchange point programmable networks," in *Advanced Information Networking and Applications: Proceedings of the 33rd International Conference on Advanced Information Networking and Applications (AINA-2019) 33*. Springer, 2020, pp. 485–497.

[4] R. Durner and W. Kellerer, "Network function offloading through classification of elephant flows," *IEEE Transactions on Network and Service Management*, vol. 17, no. 2, pp. 807–820, 2020.

[5] J. G. Gomez, V. Hernandez, and G. Ramirez-Gonzalez, "Traffic classification in ip networks through machine learning techniques in final systems," *IEEE Access*, 2023.

[6] M. Hamdan, B. Mohammed, U. Humayun, A. Abdelaziz, S. Khan, M. A. Ali, M. Imran, and M. N. Marsono, "Flow-aware elephant flow detection for software-defined networks," *IEEE Access*, vol. 8, pp. 72 585–72 597, 2020.

[7] C. Hardegen, B. Pfülb, S. Rieger, and A. Gepperth, "Predicting network flow characteristics using deep learning and real-world network traffic," *IEEE Transactions on Network and Service Management*, vol. 17, no. 4, pp. 2662–2676, 2020.

[8] C. Hardegen, B. Pfülb, S. Rieger, A. Gepperth, and S. Reißmann, "Flow-based throughput prediction using deep learning and real-world network traffic," in *2019 15th International Conference on Network and Service Management (CNSM)*. IEEE, 2019, pp. 1–9.

[9] P. Jurkiewicz, "Boundaries of flow table usage reduction algorithms based on elephant flow detection," in *2021 IFIP Networking Conference (IFIP Networking)*. IEEE, 2021, pp. 1–9.

[10] P. Jurkiewicz, G. Rzym, and P. Boryło, "Flow length and size distributions in campus internet traffic," *Computer Communications*, vol. 167, pp. 15–30, 2021.

[11] W.-X. Liu, J. Cai, Y. Wang, Q. C. Chen, and J.-Q. Zeng, "Fine-grained flow classification using deep learning for software defined data center networks," *Journal of Network and Computer Applications*, vol. 168, p. 102766, 2020.

[12] P. Megyesi and S. Molnár, "Analysis of elephant users in broadband network traffic," in *Advances in Communication Networking: 19th EUNICE/IFIP WG 6.6 International Workshop, Chemnitz, Germany, August 28-30, 2013. Proceedings 19*. Springer, 2013, pp. 37–45.

[13] P. Jurkiewicz, "flow-models: A framework for analysis and modeling of IP network flows," *SoftwareX*, vol. 17, p. 100929, 2022.

[14] P. Poupart, Z. Chen, P. Jaini, F. Fung, H. Susanto, Y. Geng, L. Chen, K. Chen, and H. Jin, "Online flow size prediction for improved network routing," in *2016 IEEE 24th International Conference on Network Protocols (ICNP)*. IEEE, 2016, pp. 1–6.

[15] A. Shaikh, J. Rexford, and K. G. Shin, "Load-sensitive routing of long-lived ip flows," *ACM SIGCOMM Computer Communication Review*, vol. 29, no. 4, pp. 215–226, 1999.

[16] G. Shen, Q. Li, S. Ai, Y. Jiang, M. Xu, and X. Jia, "How powerful switches should be deployed: A precise estimation based on queuing theory," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 811–819.

[17] P. Xiao, W. Qu, H. Qi, Y. Xu, and Z. Li, "An efficient elephant flow detection with cost-sensitive in sdn," in *2015 1st International Conference on Industrial Networks and Intelligent Systems (INISCom)*. IEEE, 2015, pp. 24–28.

[18] H. Xu, H. Huang, S. Chen, and G. Zhao, "Scalable software-defined networking through hybrid switching," in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 2017, pp. 1–9.

[2] https://pytorch.org/

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

Porównanie otwartoźródłowych implementacji sterowników SDN: OpenDaylight, Open Network Operating System, Floodlight

Comparison between open source SDN controllers: OpenDaylight, Open Network Operating System, Floodligh

Aleksandra Mardaus[1]; Jerzy Domżał[1]

amardaus@student.agh.edu.pl, jerzy.domzal@agh.edu.pl

DOI: 10.15199/59.2023.4.93

Streszczenie: Praca ma na celu porównanie wybranych otwarto-źródłowych implementacji sterowników SDN. Parametry związane z wydajnością zmierzono za pomocą programu CBench. Wykonano symulację pozwalającą zmierzyć czasy reakcji na awarię łącza i awarię sterownika. Do przeprowadzenia testów bezpieczeństwa wykorzystano dedykowane narzędzie sdnpwn. Zebrane informacje pozwoliły wskazać mocne i słabe strony poszczególnych sterowników, umożliwiając określenie odpowiednich zastosowań dla każdego z nich.

Abstract: The aim of the presented work is to compare selected open-source implementations of SDN controllers. Throughput and latency measurements were taken using the CBench program. The simulation of link failure and controller failure were made using the API provided by Mininet network simulator. For detecting security vulnerabilities, a dedicated program sdnpwn was used. The work provides an overview of the selected controllers, indicating their strengths and weaknesses.

Słowa kluczowe: SDN, sterownik, wydajność, niezawodność, bezpieczeństwo

Keywords: SDN, controller, performance, resilience, security

1. WSTĘP

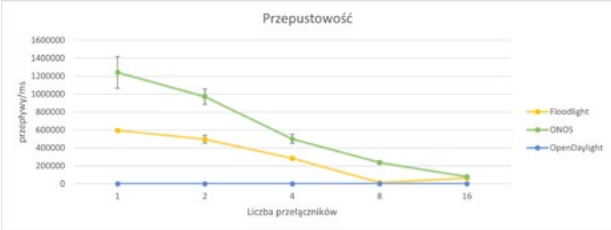
Software-Defined Networking (SDN), czyli sieci sterowane programowo to dobrze znana i ugruntowana już koncepcja podejścia do budowania sieci telekomunikacyjnych. Polega ona na oddzieleniu warstwy sterowania od warstwy danych, umożliwiając administratorom centralne zarządzanie przepływami w sieci. Zastosowanie takiego podejścia eliminuje konieczność ręcznego konfigurowania urządzeń i ułatwia wprowadzanie zmian. Pozwala także na uniezależnienie od wytwórców sprzętu i wykorzystywanie w sieci urządzeń różnych producentów. Sterownik SDN stanowi centralny punkt sieci, odpowiadając za całą inteligencję i decyzyjność w obrębie sieci. Wśród wielu różnych sterowników SDN istnieją zarówno rozwiązania komercyjne jak i o otwartym kodzie źródłowym. W pracy wybrano do porównania trzy sterowniki open-source, czyli takie, w których kod źródłowy jest publicznie dostępny, każdy może go używać, modyfikować i rozpowszechniać. Sterowniki uwzględnione w pracy to OpenDaylight [1], Floodlight [2] i Open Network Operating System [3]. Przeprowadzono eksperymenty mające na celu porównanie ich wydajności, odporności na awarie i wybranych aspektów bezpieczeństwa. Wskazano możliwości i ograniczenia każdego sterownika. Analiza pozwala na wybór najlepszego rozwiązania w zależności od potrzeb.

2. IMPLEMENTACJA

Badania zostały przeprowadzone w środowisku wirtualnym - pracę zrealizowano z użyciem hipernadzorcy VirtualBox. Utworzone zostały 3 maszyny wirtualne z systemem operacyjnym Ubuntu 22.04 LTS Jammy Jellyfish. Ponadto została utworzona maszyna wirtualna z Ubuntu 14.04, na której był uruchamiany program do testowania wydajności CBench - jego działanie wymaga bowiem użycia bibliotek nieobecnych na nowszych systemach operacyjnych. Dodatkowo utworzono maszynę z Kali Linuxem.

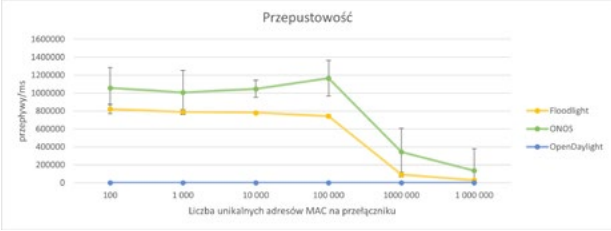
3. WYDAJNOŚĆ

Do pomiarów wydajności wykorzystano program CBench (Controller Benchmarker) [4]. Jest to dedykowane narzędzie do przeprowadzania testów wydajności sterowników wykorzystujących protokół OpenFlow [5]. CBench emuluje duże liczby przełączników, które łączą się ze sterownikiem, generując komunikaty PacketIn dla nowych przepływów. Jako metryki wydajności sterowników wykorzystuje opóźnienie i przepustowość.
Przepustowość mierzona jest jako szybkość przetwarzania przepływów przez sterownik. Zliczana jest liczba odebranych pakietów PacketOut względem liczby wysłanych pakietów PacketIn w jednostce czasu. Opóźnienie sterownika to czas, którego potrzebuje sterownik na przetworzenie pojedynczego pakietu. CBench w trybie pomiaru opóźnienia wysyła do sterownika pojedynczy pakiet PacketIn i czeka na odpowiedź przed wysłaniem kolejnego pakietu. Operacja ta jest powtarzana wielokrotnie w zadanym przedziale czasowym, a następnie obliczany jest średni czas od wysłania pakietu do otrzymania odpowiedzi od sterownika.
Uruchomiono program CBench i dokonano kolejnych pomiarów zmieniając liczbę przełączników i liczbę unikalnych adresów MAC na przełączniku. Wyniki przedstawiono na rysunku 1.



Rysunek 1: Przepustowość względem liczby przełączników.

OpenDaylight zamiast prawidłowych wyników zwrócił same zera. Problem ten jest powszechnie znany jako "all-zeros problem" i został zaobserwowany również przez autorów artykułu [6]. Ponieważ program CBench nie jest w stanie zmierzyć przepustowości sterownika OpenDaylight, został na wykresie oznaczony zerami i nie uwzględniono go w obliczeniach. W przypadku pozostałych sterowników zaobserwowano, że ONOS zapewnia dwukrotnie większą przepustowość niż Floodlight. Dla obydwu sterowników zaobserwowano proporcjonalną zależność - przepustowość maleje w sposób liniowy wraz z podwajaniem liczby przełączników. Możliwą przyczyną takiej zależności jest narzut ruchu kontrolnego pojawiający się przy zwiększaniu liczby urządzeń.

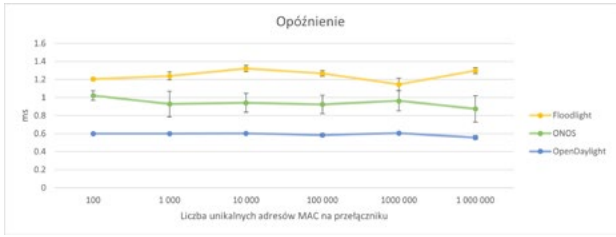


Rysunek 2: Przepustowość względem liczby adresów MAC na przełączniku.

Rysunek 2 przedstawia zmiany przepustowości względem liczby unikalnych adresów MAC na przełączniku. OpenDaylight ponownie nie zwrócił prawidłowych wyników. Zaobserwowano, że ONOS zapewnia większą przepustowość niż Floodlight. W obydwu przypadkach przepustowość nie zmienia się znacząco, gdy liczba unikalnych adresów MAC na przełączniku wynosi od 100 do 100 000. Istotny spadek przepustowości następuje po przekroczeniu 100 000 adresów na przełączniku. Może to jednak nie być spowodowane wydajnością sterownika, a ograniczeniami środowiska symulacyjnego. Pomiary związane ze sterownikiem ONOS charakteryzują się większym odchyleniem standardowym, a w konsekwencji przedziały ufności są znacznie większe niż dla Floodlight.

Rysunek 3: Opóźnienie względem liczby przełączników.

Na rysunku 3 przedstawiono zależności zmian opóźnienia od liczby przełączników. Najmniejsze opóźnienie zapewnia OpenDaylight. Jest ono w przybliżeniu funkcją stałą, gdyż wartości opóźnienia są niezależne od liczby przełączników w sieci. Kształt wykresów dla przełączników Floodlight i ONOS jest zbliżony do funkcji liniowej, rosnącej wraz ze wzrostem liczby przełączników. Floodlight uzyskuje gorsze wyniki, osiągajac większe opóźnienia.



Rysunek 4: Opóźnienie względem liczby unikalnych adresów MAC na przełączniku.

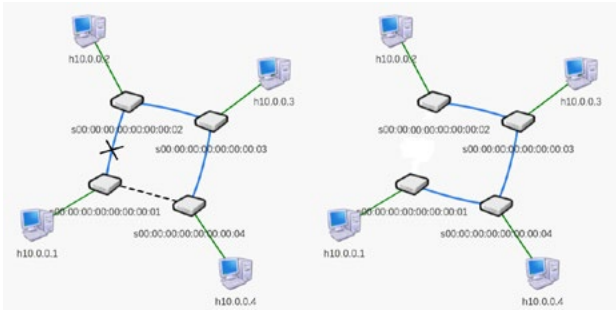
Rysunek 4 przedstawia zmiany opóźnienia względem liczby unikalnych adresów MAC na przełączniku. Nie zaobserwowano zależności między opóźnieniem a liczbą unikalnych adresów MAC na przełączniku. Biorąc pod uwagę niepewności pomiarowe widoczne na wykresie jako słupki błędów, funkcję opisującą tę zależność można porównać do funkcji liniowej o stałej wartości. Najmniejsze opóźnienia osiąga OpenDaylight, a największe Floodlight. Na podstawie uzyskanych wyników można wywnioskować, że opóźnienie nie zależy od liczby unikalnych adresów MAC na przełączniku.

4. ODPORNOŚĆ NA AWARIE

W celu porównania odporności na awarie sterowników przeprowadzono dwie symulacje: awarii łącza oraz awarii sterownika.

4.1 Awaria łącza

W celu zasymulowania awarii łącza wykorzystano skrypt napisany w języku Python wykorzystujący Mininet API. Zbudowano sieć złożoną z 4 przełączników w topologii liniowej i do każdego przełącznika przyłączono po jednym hoście (rys. 10). Dodatkowo dodano połączenie pomiędzy dwoma niepołączonymi przełącznikami, które było używane jako łącze zapasowe w przypadku awarii.



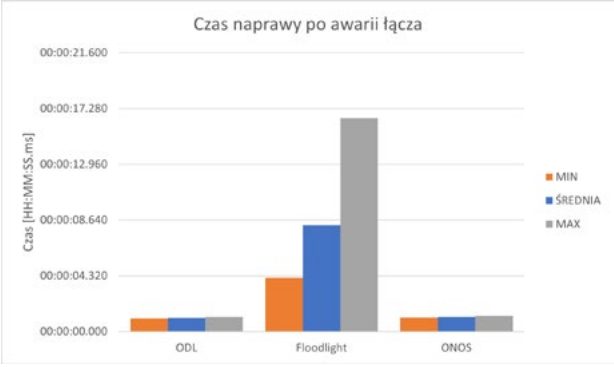
(a) Topologia przed awarią łącza.

(b) Topologia po awarii łącza.

Rysunek 5: Symulacja awarii łącza.

Czas awarii łącza zmierzono jako czas od utracenia łączności pomiędzy h1 i h2 spowodowanego wyłączeniem bezpośredniego łącza pomiędzy nimi, do odzyskania łączności po przełączeniu się na łącze zapasowe. Zaobserwowano, że OpenDaylight i ONOS potrzebują podobną ilość

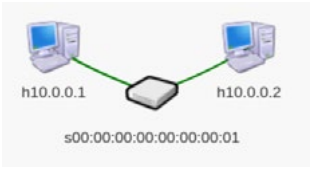
czasu na przełączenie się na trasę alternatywną w przypadku awarii łącza - około 1 sekundę. Floodlight wypadł tu gorzej, potrzebował bowiem średnio 8, a maksymalnie nawet 16 sekund do przełączenia się na nową trasę. Wyniki przedstawiono na rysunku 6.



Rysunek 6: Wykres czasu przywrócenia działania sieci po awarii łącza.

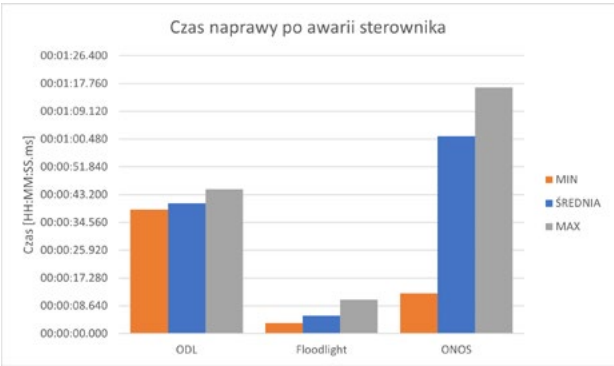
4.2 Awaria sterownika

Symulację awarii sterownika również przeprowadzono w Mininecie. Wykorzystano prostą topologię złożoną z jednego przełącznika i dwóch hostów do niego podłączonych (rys. 7). Taka topologia nie wymaga użycia zewnętrznego skryptu, dlatego w tym przypadku Mininet został uruchomiony bezpośrednio.



Rysunek 7: Topologia użyta w symulacji awarii sterownika.

Czas awarii sterownika zmierzono jako czas od utracenia łączności pomiędzy h1 i h2 spowodowanego wyłączeniem sterownika, do przywrócenia osiągalności po jego ponownym włączeniu. Podczas przeprowadzania eksperymentu zaobserwowano, że w przypadku użycia sterownika OpenDaylight, transmisja nie zostaje przerwana pomimo jego wyłączenia. Przełączniki nie usuwają wpisów z tablic pomimo utracenia łączności ze sterownikiem. Można uznać to za działanie korzystne, bowiem pomimo awarii sterownika działanie sieci jest zachowane.
Aby jednak uwzględnić OpenDaylight w porównaniu, po każdej symulowanej awarii ręcznie usuwano przepływy z przełącznika. Otrzymane wyniki przedstawiono na rys. 8.



Rysunek 8: Wykres czasu przywrócenia działania sieci po awarii sterownika.

Jak wynika z przeprowadzonych badań, w przypadku awarii łącza sterowniki OpenDaylight i ONOS potrzebują około 1 sekundę by przełączyć się na trasę alternatywną. Floodlight natomiast wymaga w tym celu średnio 8 sekund. W przypadku awarii sterownika natomiast Floodlight uzyskał najlepszy czas przywrócenia działania. Uruchamia się on znacznie szybciej niż pozostałe sterowniki – średnio 5 sekund, podczas gdy OpenDaylight potrzebuje w przybliżeniu 40 sekund, a ONOS aż minutę.

5. BEZPIECZEŃSTWO

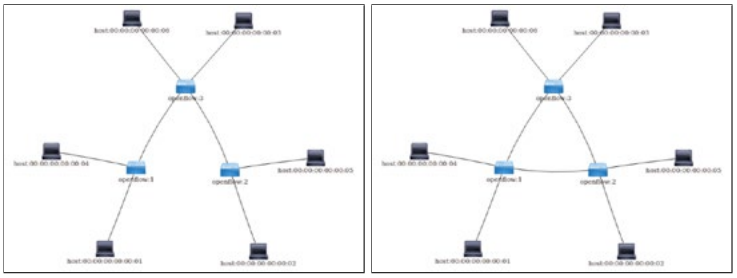
Do testów penetracyjnych wykorzystano sdnpwn, dedykowane narzędzie do badania bezpieczeństwa SDN. Rozwiązanie to umożliwia przeprowadzenie znanych ataków na wybranych sterownikach w celu rozpoznania podatności.

5.1 Wykrywanie obecności sterownika

Pierwszym przeprowadzonym eksperymentem było sprawdzenie, czy możliwe jest wykrycie i rozpoznanie sterownika w sieci. Sprawdzono to za pomocą modułu sdnpwn.py controller-detect.
Sterownik OpenDaylight został wykryty, choć jego wersja została rozpoznana nieprawidłowo. Uruchomiony ODL Nitrogen został błędnie rozpoznany jako Hydrogen - pierwsza wersja sterownika. Sprawdzono również, jak rozpoznawane są inne wersje (Lithium, Boron), za każdym razem uzyskując ten sam wynik. Może to świadczyć o braku możliwości rozponania wersji bez dostępu do sterownika lub o braku takiej funkcjonalności w narzędziu sdnpwn.
Numer portu, na którym znajduje się interfejs graficzny został prawidłowo zidentyfikowany, co może stanowić zagrożenie. Nieuprawniony użytkownik, wiedząc na którym porcie dostępny jest Graphical User Interface (GUI), może próbować się do niego zalogować. W przypadku pozostawienia przez administratora domyślnego hasła lub ustawienia hasła łatwego do złamania, atakujący będzie mógł z poziomu GUI przeprowadzać zmiany w sieci.
Floodlight został rozpoznany prawidłowo, ONOS również. W obydwu przypadkach nie została sprecyzowana wersja, co prowadzi do konkluzji, że sdnpwn w ogóle nie rozpoznaje wersji.

5.2 LLDP Replay

W sieciach SDN wykorzystywany jest protokół Link Layer Discovery Protocol (LLDP) do wykrywania łączy pomiędzy urządzeniami. Sterownik na podstawie pakietów LLDP otrzymanych od przełączników rozpoznaje topologię danej sieci. Atak LLDP Replay polega na przechwyceniu pakietu LLDP wygenerowanego przez wybrany przełącznik i wysyłaniu go z innego miejsca w sieci [7], [8]. Skutkuje to przesłaniem do sterownika wiadomości zawierającej informacje o łączu, którego w rzeczywistości nie ma w sieci. Sterownik uwzględnia to łącze w topologii i próbuje przez nie wysyłać pakiety, które wówczas nie dotrą do odbiorcy.



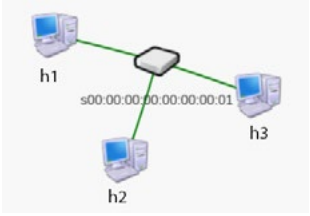
(a) Przed atakiem. (b) Po ataku.

Rysunek 9: Topologia użyta w ataku LLDP Replay.

Na rys. 9 przedstawiono topologie użyte podczas wykonywania ataku LLDP Replay. Podczas eksperymentu zaobserwowano, że ONOS jako jedyny z porównywanych sterowników posiada zabezpieczenia przed atakiem LLDP Replay. Floodlight i OpenDaylight nie weryfikują pakietów LLDP i w związku z tym mogą przechowywać nieprawidłowe informacje o topologii, co w następstwie może prowadzić do utraty pakietów, bądź przesyłania ich do niewłaściwego odbiorcy.

5.3 ARP Poisoning

ARP Poisoning, to rodzaj ataku, w którym atakujący wysyła sfałszowane wiadomości ARP do innego urządzenia w sieci. Urządzenie zamieszcza wpis w pamięci podręcznej ARP wiążący adres IP z rzeczywistego urządzenia w sieci z adresem MAC atakującego. Wówczas atakujący może odbierać wiadmości kierowane do innego urządzenia [9]. Nie jest to rodzaj ataku specyficznego dla sieci SDN, jednak centralnie zarządzana sieć powinna implementować mechanizmy zabezpieczające przed takim atakiem. Sterownik powinien monitorować cały ruch ARP w sieci i zapobiegać "zatruwaniu" go. Sposoby zapobiegania atakowi ARP Poisoning zostały opisane w [10].



Rysunek 10: Topologia wykorzystana w ataku ARP Poisoning.

Atak przeprowadzono z użyciem modułu dp-arp-poison programu sdnpwn. Pozwolił on na zastąpienie adresu MAC hosta h2 w tablicy ARP h1 adresem MAC atakującego (h3), co sprawiło, że cały ruch przeznaczony do h2 był kierowany do h3 zamiast do h2. Procedurę tę przeprowadzono kolejno dla sieci ze sterownikiem OpenDaylight, Floodlight i ONOS. W każdym przypadku atak zakończył się powodzeniem. Brak zabezpieczenia przed ARP Poisoning oznacza możliwość przeprowadzenia innych ataków sieciowych, takich jak podsłuch, Man-In-The-Middle oraz MAC Flooding. Stanowi to poważne zagrożenie dla bezpieczństwa sieci.

6. WNIOSKI

Sterownik OpenDaylight obecnie jest słabo rozwijany, co wynika z niewystarczająco licznej i zaangażowanej społeczności tworzącej go. Jego kolejne wersje znacząco się od siebie różnią w kwestii wspieranych modułów, co może być przyczyną problemu w przypadku chęci aktualizacji do nowszej wersji. W pomiarach opóźnienia uzyskuje on bardzo dobre wyniki, jednak nie współpracuje z narzędziem Cbench w trybie pomiaru przepustowości. Floodlight oferuje przeciętną wydajność, natomiast jego zaletą jest stabilne działanie i krótki czas uruchamiania. Jest najlżejszym spośród porównywanych sterowników, posiada także najmniej znanych podatności opisanych w CVE. ONOS zapewnia najlepszą wydajność, osiąga dwukrotnie większą przepustowość niż Floodlight. Wypada on jednak najgorzej jeśli chodzi o czas reakcji na awarie. Jako jedyny posiada zabezpieczenia przed LLDP Replay. Uzyskane wyniki prowadzą do wniosku, iż żadnego z porównywanych sterowników nie można nazwać obiektywnie najlepszym. Każdy posiada wady i zalety, a decyzja o wyborze rozwiązania dla danej sieci musi być podjęta przez administratora z uwzględnieniem obserwacji opisanych w pracy.

LITERATURA

[1] "Opendaylight." https://www.opendaylight.org/.

[2] "Floodlight." https://floodlight.atlassian.net/wiki/spaces/floodlightcontroller/overview.

[3] "Onos." https://opennetworking.org/onos/.

[4] "Cbench." https://floodlight.atlassian.net/wiki/spaces/floodlightcontroller/pages/1343657/Cbench+New.

[5] C. Laissaoui, N. Idboufker, R. Elassali, and K. E. Baamrani, "A measurement of the response times of various openflow/sdn controllers with cbench," *IEEE/ACS 12th AICCSA Conference*, 2015.

[6] M. M. Elmoslemany, A. S. T. Eldien, and M. M.Selim, "Performance analysis in software defined network controllers," *5th ICCES Conference*, 2020.

[7] T.-H. Nguyen and M. Yoo, "Analysis of link discovery service attacks in sdn controller," *International Conference on Information Networking (ICOIN)*, 2017.

[8] S. Hong, L. Xu, and G. G. Haopei Wang, "Poisoning network visibility in software-defined networks: New attacks and countermeasures," *NDSS Symposium*, 2015.

[9] Y. J. Mohammad Z. Masoud and I. Jannoud, "On preventing arp poisoning attack utilizing software defined network (sddn) paradigm," *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, 2015.

[10] Z. Shah and S. Cosgrove, "Mitigating arp cache poisoning attack in software-defined networking (sdn): A survey," *Electronics*, 2019.

KONFERENCJA RADIOKOMUNIKACJI
I TELEINFORMATYKI
KRiT 2023

MECHANIZM ADAPTACYJNEGO ZARZĄDZANIA CZASAMI ŻYCIA PRZEPŁYWÓW W SIECIACH SDN

MECHANISM FOR ADAPTIVE FLOW TIMEOUT MANAGEMENT IN SDN

Krystian Radamski[1]; Robert Wójcik[2]

[1] Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Kraków, kradamski@student.agh.edu.pl
[2] Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Kraków, robert.wojcik@agh.edu.pl

DOI: 10.15199/59.2023.4.94

Streszczenie: **Sieci sterowane programowo pozwalają na dynamiczne sterowanie ruchem w sieci. Do sprawnego działania konieczne jest odpowiednie zarządzanie tablicami przepływów w węzłach. W artykule proponujemy mechanizm zarządzania czasami życia przepływów, który wykorzystuje ustawienia parametrów idle timeout oraz hard timeout. W celu wyliczenia odpowiednich wartości parametrów mechanizm analizuje czas przybycia kolejnych pakietów przepływu. Dodatkowo, działanie algorytmu uwzględnia aktualny poziom wypełnienia tablicy wpisów.**

Abstract: **Software-Defined Networking allows dynamic control of network traffic. For efficient operation, proper flow table management is necessary. In this paper, we propose a mechanism to manage flow timeout values, which uses idle timeout and hard timeout parameters. To calculate the appropriate values of the parameters, the mechanism analyses the packet interarrival times. The algorithm also takes into account the current occupancy of the flow table.**

Słowa kluczowe: **Sieci Sterowane Programowo, SDN, czas życia przepływów, tablica przepływów.**

Keywords: **Software-Defined Network, SDN, flow timeout, flow table.**

1. WSTĘP

Sieci sterowane programowo (ang. SDN – Software-Defined Networks), rozdzielają warstwę przekazywania danych od warstwy sterowania. Warstwa sterowania, za którą odpowiada centralny sterownik jest w pełni programowalna, co oznacza, że urządzenia sieciowe pozwalają na przesyłanie pakietów zgodnie z regułami zdefiniowanymi przez sterownik SDN. Komunikacja pomiędzy sterownikiem a urządzeniami sieciowym wymaga reguł umożliwiających nawiązanie połączenia. Jednym z najbardziej popularnych rozwiązań w tej dziedzinie jest protokół OpenFlow.

Przełącznik zgodny ze standardem OpenFlow oparty na koncepcji tablicy przepływów (ang. flow table), w której są zamieszczane przepływy. Jeśli ruch przychodzący do przełącznika zostanie dopasowany do przepływu w tablicy, to jest on obsługiwany przez urządzenie zgodnie z zapisanymi w tablicy regułami. W przeciwnym przypadku węzeł kieruje zapytanie do centralnego sterownika. Sterownik może dodawać, aktualizować i usuwać wpisy w tablicach przepływów, zarówno reaktywnie w odpowiedzi na pakiety, jak i proaktywnie instalując regułę zanim jakikolwiek pakiet pojawi się w przełączniku. Przepływy zainstalowane w tablicy przepływów, dzięki zawartej w nich instrukcji pozwalają na podjęcie decyzji, co należy zrobić z odebranym pakietem. Przepływy mogą być wpisane do tablicy przepływów na stałe, lub czasowo. Przepływy wpisane na stałe są usuwane tylko na polecenie sterownika. Przepływy wpisane czasowo (domyślny tryb) są usuwane po przekroczeniu czasu życia przepływu. Istnieją dwa parametry określające czasu życia przepływu:

- hard timeout – wartość tego parametru oznacza, iż przepływ zostanie usunięty po zadanej liczbie sekund, niezależnie od jego aktywności;
- idle timeout – wartość tego parametru oznacza, iż przepływ zostanie usunięty po zadanej liczbie sekund braku aktywności (nowych pakietów).

Wpisy do tablicy przepływów wymagają zasobów pamięci. Tablice przepływów są powszechnie implementowane w trójdzielnych pamięciach adresowalnych TCAM (ang. ternary content-addressable memory). Z racji ogromnego ruchu występującego w sieciach i konieczności szybkiego decydowania gdzie taki ruch kierować, potrzebna jest wysoka wydajność tych pamięci. Pamięć TCAM umożliwia bardzo szybkie przeglądanie zapisanych w niej wpisów przepływów, co wiąże się jednak z wysokim kosztem produkcji oraz dużym poborem energii [3]. Optymalizacja wielkości potrzebnych tablic jest więc istotnym problemem badawczym.

2. PRZEGLĄD LITERATURY

Dotychczas przeprowadzono wiele badań w obszarze SDN związanych z zarządzaniem tablicami przepływów. Część rozwiązań skupia się na redukcji liczby przepływów w tablicy przez zastosowanie agregacji wpisów. Niektóre prace skupiały się na podziale zasad i rozprowadzaniu ich w sieci do przełączników w zależności od pojemności tych urządzeń. Spory obszar badań obejmuje rozwiązania, które dotyczą zarządzania mechanizmami pozwalającymi ustawiać odpowiednie czasy życia przepływów. Istnieją również rozwiązania, które proponują metodę buforowania reguł w dodatkowej pamięci kontrolera.

Intelligent Timeout Master [7] jest mechanizmem, który oblicza odpowiednią wartość parametru idle ti-

meout na podstawie cech przepływu. Początkowo przepływ instalowany jest z wartością inicjalizującą w tablicy przepływów i jest zapisywany z odpowiednim znacznikiem czasu w pamięci kontrolera. TimeoutX [6] stosuje zarządzanie na podstawie tego, jak długo wpis zajmuje miejsce w pamięci, następnie na tej podstawie ustawia odpowiednie dla nich adaptacyjne limity czasu. Wykorzystuje trzy parametry: szacowany czas trwania przepływu, typ przepływu oraz wskaźnik wykorzystania tablicy.

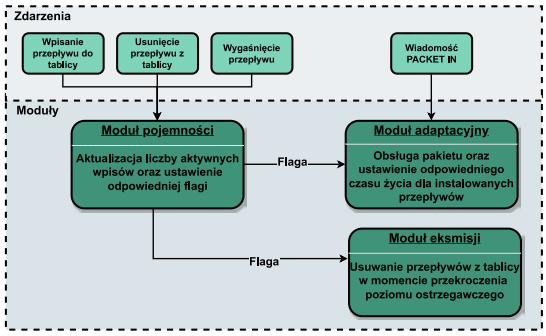
Dynamic Hard Timeout [2] ma na celu poprawienie przydziału pamięci TCAM poprzez odpowiednie ustawienie parametru hard timeout. Metoda stosuje dynamiczny przydział wartości licznika zarówno dla przewidywalnych, jak i nieprzewidywalnych przepływów. W tym celu metoda śledzi małą liczbę ostatnich pakietów przepływu, w celu zgromadzenia potrzebnych informacji do wyliczenia czasu pomiędzy przybyciami pakietów. W celu eksmisji oraz usuwania pakietów jest wykorzystywany algorytm LRU, który jest uruchamiany w momencie przekroczenia zadanego progu, w tym wypadku ustawionego na 90%.

IHTA jest metodą, która jednocześnie proponuje ustawienie zarówno czasu życia przepływów parametrem idle timeout oraz hard timeout. W celu przypisania odpowiednich czasów wyznaczany jest czas pomiędzy przybyciami pakietów. Natomiast usuwanie wpisów z tablicy odbywa się w oparciu o liczbę zliczonych pakietów *N*, które zostały dopasowane do danego przepływu w tablicy wpisów przełącznika. Metoda ta dzieli wpisy na dwie grupy: wpisy które miały dopasowanie pakietów poniżej wartości *N* oraz wpisy, które zostały dopasowane powyżej wartości *N*. W pierwszej kolejności usuwane są wpisy, które miały mniej dopasowań [4].

Podejście [1] proponuje wykorzystanie historii referencyjnej przepływów do nadawania każdemu z nich odpowiedniego stopnia ważności. Stopnie ważności są dynamicznie zmieniane i aktualizowane, wykorzystuje się je do określenia najbardziej odpowiedniego wpisu, który ma zostać usunięty z tablicy.

3. ADAPTACYJNE ZARZĄDZANIE CZASAMI ŻYCIA PRZEPŁYWÓW

Opracowany mechanizm adaptacyjnego zarządzania życiem przepływów w przełączniku OpenFlow, to mechanizm, który wprowadza metodę pozyskiwania informacji o czasach przybycia pomiędzy kolejnymi pakietami. Wdraża koncepcję stosowania jednocześnie obu parametrów odpowiedzialnych za ustawienie czasu życia przepływów. Jak dotychczas przeprowadzono tylko jedne badania w tym obszarze [4]. Wykorzystuje już znane koncepcje dotyczące zarządzania tablicą przepływów, jak mechanizm odpowiedzialny za eksmisję pakietów. Zostaje wprowadzona również koncepcja trójstrefowego podejścia do adaptacyjnego ustawienia czasów życia przepływów w zależności od aktualnego stopnia wykorzystania pamięci TCAM przełącznika. Dla sformułowanego problemu rozwiązanie zostało podzielone na trzy główne aspekty działania adaptacyjnego mechanizmu. Elementy, na które składa się rozwiązanie to moduł adaptacyjny, moduł pojemności oraz moduł eksmisji. Schemat modułów został zaprezentowany na rysunku 1.



Rys. 1. Schemat modułów mechanizmu adaptacyjnego

3.1. Moduł pojemności

Moduł ten odpowiada za monitorowanie zapełnienia tablicy wpisów przepływów w przełączniku. Wykorzystuje on pamięć pomocniczą kontrolera, w której zapisuje wszystkie przepływy, które są instalowane w przełączniku. Działanie modułu polega na budowaniu repliki tablicy przepływów w pamięci podręcznej, na podstawie której kontroler oblicza poziom wykorzystania pamięci TCAM przełącznika. Uruchamia się za każdym razem, gdy jest wysyłany wpis do tablicy przepływów. Sprawdzana jest wtedy zajętość tablicy przepływów, a następnie w zależności od tego jak bardzo zapełniona jest tablica przepływów, ustawiana jest odpowiednia flaga. Gdy sterownik zleca zainstalowanie przepływu w pamięci przełącznika, zostaje to również odnotowane w jego pamięci. Dzięki temu sterownik posiada informacje o liczbie aktywnych wpisów w pamięci poszczególnych przełączników. Głównym zadaniem tego modułu jest ustawianie flagi, od której zależy praca innych modułów. Flaga może być ustawiana na trzech poziomach wyznaczonych na podstawie aktualnego zapełnienia pamięci przełącznika: normalny (0-50%), wzmożonego zapełnienia (51-97%) i ostrzegawczy (>97%).

3.2. Moduł adaptacyjny

Moduł odpowiada za ustawianie czasów życia przepływów, które będą instalowane w przełączniku. Mechanizm używa zarówno czasu idle timeout oraz hard timeout. Do wyliczenia odpowiednich wartości obu parametrów wykorzystywane są informacje o charakterystyce ruchu danego przepływu oraz aktualnego zapełnienia pamięci TCAM przełącznika. Działanie modułu można podzielić na trzy etapy: inicjalizacja, ustawienie idle timeout oraz ustawienie hard timeout.

Inicjalizacja. Gdy pakiet dociera do przełącznika i nie znajduje odpowiedniego dopasowania do wpisu, wysyłana jest do sterownika wiadomość. W sterowniku każdy pakiet jest odnotowywany. Jeżeli jest to pierwszy pakiet danego przepływu zostaje on obsłużony właśnie przez moduł inicjalizujący.

Idle timeout. Ustawienie tego parametru odbywa się dla każdego ponownego zgłoszenia obsługi przepływu, który wystąpił wcześniej. Główne obliczenia wartości czasu życia wykorzystują czas przybycia pomiędzy kolejnymi pakietami. Dodatkowo brana jest pod uwagę wartość flagi zwracana przez moduł pojemności, od

której zależy jak bardzo rygorystycznie będzie wyliczany czas życia. W celu efektywnego ustawiania parametru idle timeout w rozwiązaniu dla podstawowej wersji zastosowano wyliczanie czasu życia przepływu zgodnie ze wzorem (1).

$$t_{idle} = \max\{[r + T], t_{idle_min}\} \quad (1)$$

gdzie *r* to czas przybycia pomiędzy kolejnymi pakietami, *T* jest stałą używaną do minimalizacji wpływu potencjalnej fluktuacji odstępu między kolejnymi pakietami (najczęściej wartość *T*=1), t_{idle_min} to wartość minimalna, jaką może przyjąć parametr idle timeout (najczęściej t_{idle_min}=1).

Hard timeout. Ustawienie tego parametru pełni funkcje zabezpieczające. Cechy ruchu w centrach danych opartych o sieć SDN [5] wykazały, że średni czas trwania transferu 80% przepływów nie przekracza jedenastu sekund. Również co do górniej granicy limitu czasu życia przepływów odnosi się [2], gdzie zostało zbadane, że blisko 90% przepływów znajduję się poniżej jedenastu sekund. Badana przeprowadzone w [7] wykazały podobną zależność dla 88% przepływów. Podobnie zbadano w [8], gdzie podano, iż 90% przepływów ma czas, życia krótszy niż jedenaście sekund. Biorąc pod uwagę wszystkie przedstawione badania dotyczące trwania czasu parametr hard timeout został ustawiony na wartość równą t_{hard}=11 sekund, dla pracy przy poziomie normalnym zapełnienia tablicy przepływów. Dla poziomu wzmożonego oraz ostrzegawczego wartość parametru t_{hard} przyjmuje odpowiednio 8 i 4 sekundy.

3.3. Moduł eksmisji

Mechanizm eksmisji przepływów uruchamiany jest na podstawie wartości flagi, która ustawiana jest przez moduł pojemności. Eksmisja pakietów uruchamia się w sytuacji, gdy zapełnienie tablicy przepływów wzrośnie powyżej 97% jej rozmiaru. Wartość ta została ustalona poniżej 100%, ponieważ usuwanie wpisów w sytuacji całkowitego zapełnienia tablicy przepływów okazuje się nieefektywne. Adaptacyjny mechanizm stara się wcześniej przygotować na okoliczność zapełnienia tablicy, aby móc lepiej obsługiwać kolejne pakiety, które będą wymagały od kontrolera obsłużenia i wpisania reguły do tablicy przepływów. Mechanizm umożliwia działanie w oparciu o dowolny algorytm usuwający pakiety. W rozwiązaniu zastosowano algorytm losowego usuwania Random.

4. SYMULACJE

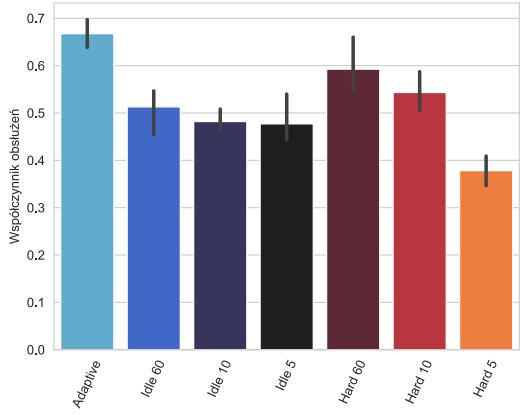
Rozwiązanie zostało zaimplementowane z wykorzystaniem sterownika RYU. Symulacje zostały wykonane w środowisku Mininet, gdzie wykorzystano przełącznik Open Vswitch. Do symulacji użyto zestawu danych pochodzącego ze zbioru grupy roboczej WIDE MAWI Working Group. Mechanizm adaptacyjnego zarządzania czasami życia przepływów porównano ze statycznymi ustawieniami parametrów idle timeout oraz hard timeout. Dla mechanizmu adaptacyjnego symulacje wykonywano z wartością *N*=3, mówiącą dla ilu pakietów należy policzyć średnią wartość czasu pomiędzy kolejnymi przybyciami pakietów. Pomiarów dokonano dla 3000 wpisów pojemności pamięci TCAM. Na wykresach zaznaczono 95% przedziały ufności wyznaczone na podstawie metody samowspornej.

4.1. Współczynnik obsłużeń pakietów

W celu porównania rozwiązania użyto współczynnika obsłużeń pakietów. Informuje on o tym, jaka liczba pakietów została dopasowana w tablicy przepływów przełącznika do liczby pakietów obsłużonych przez przełącznik i tych, które musiał obsłużyć kontroler, instalując nowy przepływ. Zdefiniowany został w następujący sposób: liczba pakietów, która przechodzi przez przełącznik z wykorzystaniem reguł zapisanych w tablicy przepływów wynosi *c*. Liczba wysłanych przez kontroler wpisów do przełącznika, modyfikujących tablicę przepływów w taki sposób, aby pakiety mogły być obsługiwane przez przełącznik wynosi *sc*. Współczynnik obsłużeń opisany jest wzorem (2).

$$Współczynnik\ obsłużeń = \frac{c}{c+sc} \quad (2)$$

Na rysunku 2 przedstawiono wyniki przeprowadzone dla symulacji z przyjętą pamięcią TCAM 3000 wpisów.



Rys. 2. Współczynnik obsłużeń

Z rysunku 2 wynika, że algorytm obsługuję te samą ilość ruchu z mniejszą liczbą zleceń wpisania reguł do pamięci przełącznika.

4.2. Wskaźnik braku pojemności

Wskaźnik braku pojemności (3) wskazuje, ile pakietów zostało niedopasowanych z powodu braku miejsca w tablicy przepływów przełącznika.

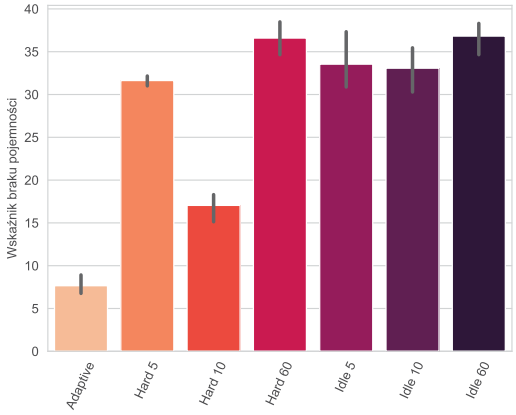
$$Wskaźnik\ braku\ pojemności = \frac{fm}{tm} * 100 \quad (3)$$

gdzie *fm* to pakiety, które nie są obsługiwane z powodu braku miejsca na nowe wpisy w przełączniku, *tm* to wszystkie pakiety, które pojawiają się w przełączniku i nie znajdują dopasowania.

Rysunek 3 przedstawia wskaźnik braku pojemności dla rozmiaru pamięci 3000 tysięcy wpisów. Oznacza to, że algorytm znacznie rzadziej obciąża całkowicie pamięć przełącznika tym samym mniejsza liczba reguł zostaje odrzucona.

4.1. Zależność parametru *N*

Proponowany mechanizm do wyliczenia średniego czasu przybycia pomiędzy kolejnymi pakietami *r* wykorzystuje *N* przyjść pakietów do sterownika. Dopiero po wystąpieniu *N* pakietów w sterowniku instalowane są przepływy z wyliczonymi wartościami czasów życia.



Rys. 3. Wskaźnik braku pojemności

Na rysunku 4 zaobserwować można niemal liniowy wzrost liczby wiadomości *PACKETIN* dla *N* od 2 do 10. Dla większych wartości liczby *N*, liczba wiadomości *PACKETIN* przychodzących do sterownika rośnie coraz wolniej.

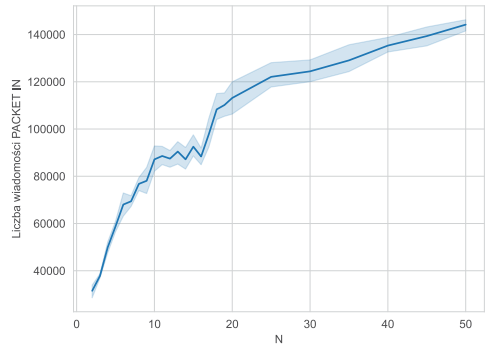
4.2. Wpływ mechanizmu na ruch w sieci

W celu sprawdzenia wpływu adaptacyjnego mechanizmu zarządzania czasami przepływów na ruch w sieci wykonano pomiar czasu obsługi dla zbioru danych. Wykorzystano funkcjonalność TCP replay, która pozwala na instalowanie przechwyconego ruchu w sieci z dowolną szybkością. Oznacza to, że ruch reprodukowany podczas symulacji wykorzystywał maksymalnie zasoby przełącznika i sterownika w taki sposób, że ruch przychodzący do przełącznika z innych hostów dopiero mógł zostać obsłużony po obsłużeniu ruchu pochodzącego ze zbioru danych.

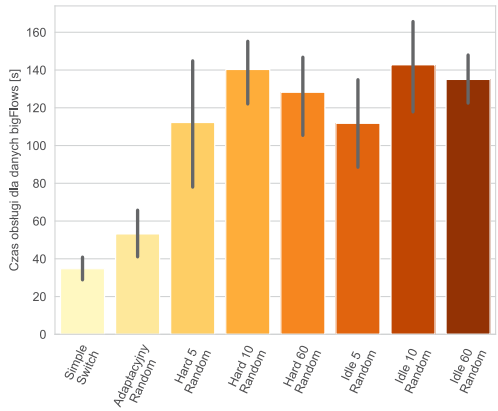
Jak wynika z rysunku 5, najszybciej ruch był obsługiwany z wykorzystaniem prostej wersji kontrolera, jednak należy mieć na uwadze, że kontroler ten nie wspiera protokołów takich jak na przykład UDP lub TCP. W porównaniu do podstawowej wersji kontrolera, adaptacyjny mechanizm działa wolniej, jednak jest szybszy w porównaniu ze statycznym ustawieniem czasów.

5. WNIOSKI

Mechanizm adaptacyjnego zarządzania czasami życia przepływów wykorzystuje zarówno ustawienie czasów życia parametrem idle timeout oraz hard timeout. Do wyliczenia wartości czasu idle timeout opracowano metodę, która umożliwia sterownikowi pozyskiwanie średniej wartości czasu pomiędzy *N* kolejnymi pakietami przepływu. Wartości czasów życia ustawiane są w zależności od aktualnego poziomu wypełnienia tablicy przepływów przełącznika. Wyniki pokazują, że prezentowany mechanizm pozwala obsłużyć te samą ilość ruchu, wykonując mniej zleceń wpisania reguł do pamięci przełącznika w porównaniu ze statycznym ustawieniem czasu życia. Dodatkowo dla prezentowanego mechanizmu mniejsza liczba pakietów jest niedopasowana z powodu braku miejsca w pamięci przełącznika, co jest ogromną zaletą proponowanego rozwiązania.



Rys. 4. Liczba generowanych wiadomości PACKETIN w zależności od N



Rys. 5. Czas obsługi wszystkich pakietów dla danego zbioru danych, przy maksymalnej szybkości symulacji

LITERATURA

[1] Abbasi M., Maleki, S., Jeon, G., Khosravi, M.R., Abdoli, H.: An intelligent method for reducing the overhead of analysing big data flows in openflow switch. IET Commun. 16, 548–559 (2022).

[2] Abinas Panda, Siddharth Shankar Samal, Ashok Kumar Turuk, Aliva Panda, Vummadi Chetty VenkateshPanda „Dynamic Hard Timeout based Flow Table Management in Openflow enabled SDN" in International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN) 30-31 March 2019

[3] Ali Ghiasian "Impact of TCAM size on power efficiency in a network of OpenFlow switches" IET Networks, pp. 367-371, November 2020

[4] Isyaku, B., et al.: IHTA: Dynamic idle-hard timeout allocation algorithm based OpenFlow switch. In: 2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE), Malaysia (2020)

[5] Lu, M.; Deng, W.; Shi, Y. TF-Idle Timeout: Improving Efficiency of TCAM in SDN by Dynamically Adjusting Flow Entry Lifecycle. In Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, Hungary, 9–12 October 2016; pp. 2681–2686

[6] Zhang, L.; Wang, S.; Xu, S.; Lin, R.; Yu, H. TimeoutX: An Adaptive Flow Table Management Method in Software Defined Networks. In Proceedings of the 2015 IEEE Global Communications Conference (GLOBECOM), San Diego, CA, USA, 6–10 December 2015; pp. 1–6

[7] Zhu, H., et al.: Intelligent timeout master: Dynamic timeout for SDN based data centers. In:2015 IFIP/IEEE International Symposium on Integrated Network Management (IM), Ottawa, Canada (2015)

Pamiętaj o prenumeracie 2023

www.elektronika.orf.pl
www.sigma-not.pl

Elektronika
konstrukcje • technologie • zastosowania

WYDAWNICTWO *SIGMA-NOT*

Instytut Telekomunikacji AGH

https://telekomunikacja.agh.edu.pl

Instytut Telekomunikacji AGH, należy do Wydziału Informatyki, Elektroniki i Telekomunikacji AGH, jest ważnym ośrodkiem kształcenia i badań naukowych w dziedzinie technik telekomunikacyjnych, ze szczególnym uwzględnieniem sieci i usług dużych szybkości. Instytut zatrudnia na pełnych etatach przeszło 60 osób; z Instytutem stale współpracuje ok. 25 doktorantów.

Od 1996 roku **Instytut Telekomunikacji AGH** uczestniczy w rozmaitych projektach europejskich, między innymi w ramach 4., 5., 6. oraz 7. Programu Ramowego. Instytut uczestniczył między innymi w opracowywaniu strategii rozwoju sieci czołowych polskich operatorów telekomunikacyjnych. Pracownicy Instytutu są konsultantami operatorów, producentów sprzętu i oprogramowania, a także instytucji regulacyjnych w Polsce i za granicą, są regularnie zapraszania do recenzowania projektów europejskich. Pracownicy Instytutu Telekomunikacji AGH działają aktywnie w stowarzyszeniach zawodowych (np. IEEE, IFIP, SEP), organizując międzynarodowe konferencje i warsztaty naukowe, a także uczestniczyli lub uczestniczą w redagowaniu renomowanych czasopism takich, jak: IEEE Communications Magazine, IEEE Transactions on Communications, Computer Communications czy Security and Communication Networks.

Instytut Telekomunikacji AGH kształci studentów na kierunkach: Teleinformatyka, Cyberbezpieczeństwo, Elektronika i Telekomunikacja oraz Electronics and Telecommunications (studia w j. angielskim).

Grupy badawcze

- Zespół Jakości Postrzeganej przez Użytkowników (QoE)
- Zespół Jakości Obsługi i Bezpieczeństwa Sieciowego (QoS & Security)
- Zespół Komunikacji Bezprzewodowej (Wireless Communications)
- Zespół Sieci Sterowanych Programowo i Wirtualizacji Sieci (SDN & NFV)
- Zespół Cyfrowego Przetwarzania Sygnałów i Sztucznej Inteligencji (DSP & AI)
- Zespół Multimediów (Multimedia)

Instytut Telekomunikacji
Wydział Informatyki, Elektroniki i Telekomunikacji AGH
ul. Czarnowiejska 74
30-059 Kraków

Tel.: +48 12 617 3937
E-mail: telekomunikacja@agh.edu.pl