IEEE COMMUNICATIONS MAGAZINE

October 2022, vol. 60, no. 10

Series: Mobile Communications and Networks

IEEE ComSoc IEEE Communications Society

A Publication of the IEEE Communications Society

www.comsoc.org

Publish your work

in the *IEEE Open Journal of the Communications Society* (OJ-COMS), the premier open access journal in communications technology!

The *IEEE Open Journal of the Communications Society* (OJ-COMS) is a fully open-access, all-electronic journal that publishes original high-quality, peer-reviewed manuscripts on advances in telecommunications systems and networks.

OJ-COMS reaches more than **5 million people** and offers a rapid review process. Submissions reporting new theoretical findings and practical contributions are welcome. Additionally, review and survey articles are considered.

Submit your work today!

www.comsoc.org/ojcoms

IEEE ComSoc | IEEE

IEEE COMMUNICATIONS MAGAZINE

October 2022, vol. 60, no. 10
www.comsoc.org/commag

IEEE COMMUNICATIONS MAGAZINE (ISSN 0163-6804) is published monthly by The Institute of Electrical and Electronics Engineers, Inc. Headquarters address: IEEE, 3 Park Avenue, 17th Floor, New York, NY 10016-5997, USA; tel: +1 (212) 705-8900; http://www.comsoc.org/commag. Responsibility for the contents rests upon authors of signed articles and not the IEEE or its members. Unless otherwise specified, the IEEE neither endorses nor sanctions any positions or actions espoused in *IEEE Communications Magazine*.

ANNUAL SUBSCRIPTION: US$71: print, digital, and electronic. US$33: digital and electronic. US$1,145: non-member print.

EDITORIAL CORRESPONDENCE: Please send e-mail to: mceditorinchief@comsoc.org.

POSTMASTER: Send address changes to *IEEE Communications Magazine*, IEEE, 445 Hoes Lane, Piscataway, NJ 08855-1331. GST Registration No. 125634188. Printed in USA. Periodicals postage paid at New York, NY and at additional mailing offices. Canadian Post International Publications Mail (Canadian Distribution) Sales Agreement No. 40030962. Return undeliverable Canadian addresses to: Frontier, PO Box 1051, 1031 Helena Street, Fort Eire, ON L2A 6C7.

SUBSCRIPTIONS: Orders, address changes — IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08855-1331, USA; tel: +1 (732) 981-0060; e-mail: address.change@ieee.org.

ADVERTISING: Advertising is accepted at the discretion of the publisher. Address correspondence to: Advertising Manager, *IEEE Communications Magazine*, IEEE, 445 Hoes Lane, Piscataway, NJ 08855-1331.

This issue contains paid advertising for information services or products.

SUBMISSIONS: Submission instructions can be found at the the magazine website: https://www.comsoc.org/publications/magazines/ieee-communications-magazine

LETTERS TO THE EDITOR: Letters to the Editor can be sent through Manuscript Central as manuscript type: "Columns (restricted)" and title "Letter to the Editor"

OMBUDSMAN: The Ombudsman shall be the first point of contact for reporting a dispute or complaint related to Society activities and/or volunteers. The Ombudsman will investigate, provide direction to the appropriate IEEE resources if necessary, and/or otherwise help settle these disputes at an appropriate level within the Society: https://www.comsoc.org/about/board-governors/ombudsman

BPA WORLDWIDE

Call for Feature Topic Proposals

IEEE COMMUNICATIONS MAGAZINE

The *IEEE Communications Magazine (ComMag)* is soliciting creative Feature Topic (FT) proposals to complement its Series and Open Call publication tracks. A Feature Topic (FT) is a group of papers focusing on a timely narrow-scope topic of interest to the readership of the magazine. They must not duplicate other publication tracks and must offer opportunities for manuscript streams not traditionally submitted to Series and Open Calls.

Potential Guest Editors (GEs) interested in organizing an FT can submit a proposal to the Editor-in-Chief (EiC) of the *IEEE Communications Magazine*. Proposers must read and fully understand the *IEEE Communications Magazine* guidelines for authors, reviewers, and editors (https://www.comsoc.org/publications/magazines/ieee-communications-magazine), and must pay special attention the **Policy for Proposing and Organizing a Feature Topic** (https://www.comsoc.org/publications/magazines/ieee-communications-magazine/policy-for-proposing-and-organizing-feature-topics). Proposals not taking all the magazine guidelines into full consideration cannot be accepted. **It is highly recommended that the lead GE communicate his topic, plan, proposed GE list to ComMag (EiC or AEiC) to discuss them before taking the time to prepare a complete proposal and/or finalizing a committed list of GEs for this matter**.

We are interested in all topics within the magazine scope and as described in the above mentioned guidelines. The preferred topics however change over time. Current topics of interest include, but are not limited to, the following:

- Communications for data centers
- Telework
- Telco IT
- Telco big data
- Mobile money
- Metaverse, virtual, augmented, and mixed reality
- Telecommunications business, economy, policies, and strategy
- Quantum communications
- Web3; Blockchain in communications and networking
- Communication technologies to mitigate climate change
- Internet neutrality and regulation
- Telecommunications infrastructures

Feature topics are not supposed to aim at attracting the type of papers that is entirely within the scope of *IEEE Communications Magazine* Series and is already submitted to them. Creative feature topic proposals that add new, innovative, and untraditional ideas to the magazine are desirable. While the topic may partially overlap with an existing Series, it must pose a unique angle that result in submission of papers that would not be traditionally submitted to this Series (quantitatively and qualitatively). For example, a cutting-edge niche topic focusing on a recent discovery/technology and its special applications can be of interest. Technical topics at the intersection with economic, social, regulatory, and legal issues/challenges are of interest. Editorial creativity and innovation are keys to having a successful proposal and feature topic.

Expand Your Communications Technology Expertise with Live, Online Training!

IEEE ComSoc Training offers high-quality, online courses led by industry experts to quickly deliver specific skills and knowledge to help you advance professionally in the communications technology field. Plus, you can earn IEEE Continuing Education Units (CEUs) with each course.

Principles of Satellite Location and Navigation

26–27 October 2022
9:00 am–12:30 pm EDT

Registration closes: 24 October at 5:00 pm EDT

Learn how satellite navigation works, what its features and limitations are, which satellite navigation systems are available or being developed and how to more effectively utilize them.

OTFS and Delay-Doppler Communications

16–17 November 2022
2:00 pm–6:00 pm EST

Registration closes: 14 November at 5:00 pm EST

Learn about delay-Doppler communications, from the fundamental theory of the Zak transform to MATLAB code functions and software-defined radio implementation.

Group Training

Is your team or organization interested in participating as a group in a ComSoc Training course? Contact Tara McNally at **t.mcnally@comsoc.org** for more information.

Get a Discount

Did you know, with an IEEE ComSoc membership you get the best price when you register for one of our courses? Not a member? Visit **www.comsoc.org/join** to sign up.

Visit www.comsoc.org/training to learn more or register today!

IEEE COMSOC TRAINING

COMSOC EMERGING TECHNOLOGIES

The IEEE Communications Society (ComSoc) Emerging Technologies Committee (ETC) is responsible for identifying and nurturing new technology directions through various activities including the formation of initiatives in areas that are of high interest to ComSoc members. In this issue of the President's Page, I am pleased to introduce Zhisheng Niu, the Chair of ETC for 2022–2023 to share with you the ongoing activities.

Xuemin (Sherman) Shen

Zhisheng Niu

ZHISHENG NIU graduated from Beijing Jiaotong University, China, in 1985, and got his M.E. and D.E. degrees from Toyohashi University of Technology, Japan, in 1989 and 1992, respectively. During 1992–1994, he worked for Fujitsu Laboratories Ltd., Japan, and in 1994 joined with Tsinghua University, Beijing, China, where he is now a professor at the Department of Electronic Engineering. During 1997–1998, he visited Hitachi Central Research Laboratory as a HIVIPS senior researcher. His major research interests include queueing theory and traffic engineering, wireless communications and mobile Internet, vehicular communications and smart networking, and green communication and networks. He has been serving IEEE Communications Society since 2000, first as Chair of Beijing Chapter (2000–2008) and then as Director of Asia-Pacific Board (2008–2009), Director for Conference Publications (2010–2011), Chair of Emerging Technologies Committee (2014–2015), Director for Online Contents (2018–2019), and currently Chair of Emerging Technologies Committee (2022–2023). He has also served as editor of *IEEE Wireless Communications* (2009–2013) and associate Editor-in-Chief of IEEE/CIC joint publication *China Communications* (2012–2016), and Editor-in-Chief of *IEEE Trans. Green Commun. & Networks* (2020–2022). He received the Outstanding Young Researcher Award from Natural Science Foundation of China in 2009, Best Paper Awards from IEEE Communication Society Asia-Pacific Board in 2013 and from *Journal of Communications and Information Networks (JCIN)* in 2019, Distinguished Technical Achievement Recognition Award from IEEE Communications Society Green Communications and Computing Technical Committee in 2018, and Harold Sobol Award for Exemplary Service to Meetings & Conferences from IEEE Communication Society in 2019. He was selected as a distinguished lecturer of IEEE Communication Society (2012–2015) as well as IEEE Vehicular Technologies Society (2014–2018). He is a fellow of both IEEE and IEICE.

It is my great pleasure to showcase the activities of the ETC which was formed in 2006 to identify and nurture new technology directions in the broad field of communications and networking areas. As communications and networking become more pervasive and interdisciplinary, the ETC's goal is to bring these cutting-edge technologies under the purview of ComSoc. Members with a common interest in a new technology are strongly encouraged to form an Emerging Technical Initiatives (ETIs) so that the new technology can be promoted by organizing a wide range of activities such as seminars, workshops, distinguished lecturer tours, etc. The ETIs may also serve as technical cosponsors for ComSoc conferences and publications, or organize a special e-issue of JSAC to showcase the emerging technologies.

In the following, I will highlight the 8 current ETIs within the ETC. More detailed information, in particular the most recent activities of each ETI, can be found from https://www.comsoc.org/about/committees/emerging-technologies-initiatives.

1. Backhaul/Fronthaul Networking & Communications

Founded in 2015, this ETI create a forum for researchers, developers and practitioners from both academia and industry to identify and discuss the backhaul/fronthaul requirements, challenges, recent development and smart end-to-end solutions pertaining to fifth-generation (5G) and beyond of mobile communication networks. It is anticipated that future networks will evolve from today's separate and incompatible fronthaul and backhaul into an integrated flexible smart wireless backhauling/fronthauling infrastructure that will support future cellular and ad hoc networks e.g., 4G/5G and 6G, Wi-Fi, IoT and emerging technologies such as driverless cars, autonomous vehicles or flying platforms, robotic control, smart buildings, and remote condition monitoring networks, etc.

2. Quantum Communications & Information Technology

Founded in 2015, this ETI aims at fostering engineering in the newly upcoming quantum technology by applying ComSoc's technical knowledge in areas like RF technology, coding theory, communications and information theory, photonic communications technology, interconnection and complexity theory, error correction, control instrumentation, modeling and simulation, communication systems architecture and hardware, optimized algorithms and applications, which all are highly required to drive quantum technology forward and get it ready for applications.

3. Network Intelligence

Founded in 2017, this ETI is to support and endorse researches to embed Artificial Intelligence (AI) in future networks, which will provide greater level of automation and adaptiveness, enabling faster deployment (from months down to minutes), dynamic provisioning adapted to the nature of network functions, and end-to-end orchestration for coherent deployment of IT and network infrastructures and service chains. It will also result in higher resiliency and better availability of future networks and services.

4. Machine Learning for Communications

Founded in 2018, this ETI is to foster research and innovation surrounding the use of machine learning (ML) for the physical (PHY) and medium access control (MAC) layers for all types of communication systems, such as wireless, optical, satellite, and molecular. It also works on establishing common data sets and related benchmarks and inviting authors to open-source their code for reproducible research.

5. Aerial Communications

Founded in 2019, this ETI focus on aerial communications from all different perspectives (vehicles, electronics, communications & networking, services), including aspects related to both aerial users and networks, as well as computational resource migra-

tion towards "portable" and flying platforms. One of the main considered domains is public safety, i.e., aerial communications with drones or balloons to deliver additional cellular coverage (for example during emergency situations) or to help first responders while providing advanced services.

6. Reconfigurable Intelligent Surfaces (RIS)
Founded in 2020, this ETI explores and supports a wide variety of research directions and standardization opportunities that rely on RIS as groundbreaking technology that will have the potential of fundamentally changing how generic wireless networks are designed and optimized today and is expected to have a big impact on the future (6G and beyond) communication solutions' design and implementation. Such a new concept has been recently proposed for a variety of applications, ranging from secure communications, non-orthogonal multiple access, millimeter-wave and terahertz communications, vehicular/aerial communications, and over-the-air-computation to improving the energy efficiency and capacity of wireless networks.

7. Integrated Sensing and Communications (ISAC)
Founded in 2021, this ETI is to explore and support a wide variety of research directions and standardization opportunities related to Integrated Sensing and Communication (ISAC), including vehicular network, environmental monitoring, remote sensing, IoT, smart city as well as indoor services such as human activity and gesture recognition. More importantly, ISAC has been recently identified as an enabling technology for B5G/6G, and the next-generation Wi-Fi system.

8. Next Generation Multiple Access (NGMA)
Founded in 2022, this ETI provides a research and networking platform for researchers to collaborate, exchange ideas, and promote initiatives on Next Generation Multiple Access (NGMA) in wireless networks. The scope of possible candidates for this ETI includes: (1) nonorthogonal multiple access in both power and code domain as well as its joint design with large-scale antenna systems; (2) massive grant-free access schemes, including random access schemes and the related signal processing techniques; (3) efficient bidirectional schemes for multiple access that include transmission of control information and data in the downlink, in addition to the uplink; (4) native AI-enabled multiple access schemes, and other possible multiple access candidates for new application scenarios such as large scale distributed machine learning; (5) other emerging multiple access schemes.

Emerging technologies are generally unknown, unproven, and risky, and therefore difficult to manage. As a result, IT organizations like ComSoc are facing with the task of not only identifying relevant emerging technologies, but also developing their organizational awareness and motivation to nurture them. In this context, the ETC has been working with the TC Restructuring Ad Hoc Committee to determine a new framework for TCs/ICs/ETIs and the migration approaches. We are also working closely with ComSoc's TE&I (Technology Evolution & Initiatives) Committee to strengthen the tie with external initiatives, i.e. those originated from IEEE-level or multi-Society initiatives. As the first trial, we will co-organize a session with TE&I at IEEE Globecom 2022 in Rio de Janeiro, Brazil, to introduce TE&I and ETIs activities to broader members.

Looking forward, we would like to encourage all ComSoc members to widen your eyes across all the potential fields linked directly or indirectly to the discipline of communications and to propose new ETIs accordingly. We also encourage all readers to contact us if you have other ideas about how ComSoc can promote and participate in emerging technologies, which will further enhance its leadership and vision in the field of communications and networking. Thank you.

CRUCIBLE OF COMMUNICATIONS: HOW AMATEUR RADIO LAUNCHED THE INFORMATION AGE AND BROUGHT HIGH TECH TO LIFE

PART 1: THE BIRTH AND BREADTH OF THE HAM RADIO HOBBY

INVITED ARTICLE

Theodore. S. Rappaport, N9NB, IEEE Fellow, Lee/Weber Chaired Professor of Electrical and Computer Engineering, Founding Director of NYU WIRELESS, New York University

ABSTRACT

The hobby of amateur radio, or "ham radio" as it is commonly known among its 3 million global practitioners, has been at the vanguard of electrical and computer engineering since a young Italian inventor, Guglielmo Marconi, first demonstrated wireless at his summer home in Bologna in 1894. Ever since that fateful discovery, amateur radio has played vast and vital roles in capturing the imagination of inventors, spawning new technologies, fueling the global engineering work force, and fostering friendship and international goodwill. This three-part series of articles chronicles the historical evolution of amateur radio, and shows the astounding impact that the ham radio hobby has made on both the engineering profession and the world.

INTRODUCTION

Since the beginning of the 20th century, the hobby of amateur radio (or "ham radio" as it is affectionately known) has incubated a global arsenal of trained technical experts, and has served as the world's proving ground for wireless communications technologies. The hobby has also provided a unique social melting pot for people from all walks of life — from ages 5 to 109 — to share in the passion of wireless communications and experimentation [1, 2]. The avocation of ham radio allows one to tinker with communication systems, software, electronics, and antennas, facilitating a very strong intuitive understanding of fundamental principles of science, technology, engineering, and math (STEM). At the same time, practitioners of ham radio develop social skills and self-confidence by sharing and learning their craft with others, and gain access to a global network of friends, colleagues, and mentors through their pursuits of the hobby.

Using tiny slivers of licensed radio frequency bands throughout the entire electromagnetic spectrum, amateur radio operators around the world are granted access to the airwaves through the Radio Regulations of the International Telecommunication Union — Radiocommunication Standardization Sector (ITU-R) [3]. Ham operators are licensed by their own country's government, assigned a unique call sign identifier for communicating throughout the world, and are permitted to operate their own hobby radio stations for the purposes of two-way communications, experimentation, and enjoyment. More than a dozen shared ham bands are available and harmonized across the globe, from the lowest carrier frequency of 136 kHz (the 2200 meter band, with a 1 W power limit) to the highest band of 47 GHz in the millimeter-wave regime.

The term "ham" is believed to have originated from wireline telegraph operators who used the word to denigrate other operators who did not receive Morse code accurately, or who had poor or unintelligible sending styles (e.g., bad fists). In G.M. Dodge's 1903 book, *The Telegraph Instructor*, a "ham" is described as a poor operator, a "plug." It seems that early professional wireless operators brought that pejorative from their wireline practice to describe the non-paid hobbyists who were building their own radio stations and were causing radio interference, but the hobby embraced the label from the onset.

Hams design, build, or buy a wide range of transmitters, receivers, amplifiers, antennas, audio and radio frequency filters, computer controllers, switching systems, and other gadgets in a constant quest to improve their stations or to enhance operations for a particular facet of the hobby that most intrigues them. Using their personal radio stations, which they call "rigs," hams engage in experimentation and discovery, enjoy leisurely conversations with other hams, try out new modes of digital and analog communications, and pursue personal interests among the many varied aspects of the hobby, such as long-distance (DX) communications, contesting, emergency preparedness, moon bounce, satellite communications, model airplane remote control, and county hunting, just to name a few. Many ham operators have "the gift of gab" and enjoy the art of "rag chewing," a good-natured term for describing an extremely long conversation with other hams over the air.

Each country provides its own licensing structure to allow citizens to gain their amateur radio license, offering different classes of license that incentivize and reward applicants to attain greater demonstrated levels of technical knowledge and operating proficiency in exchange for greater access to the amateur radio spectrum. Maximum station transmitter power is limited to about 1 kW, depending on country, the class of license, and particular frequency band. As part of the ITU spectrum allocation, there is international agreement that amateur radio frequencies are provided strictly for hobby use, and must be open for anyone to intercept and eavesdrop over the air to promote self-policing, although some modulation and coding methods make eavesdropping more technically difficult. Using ham radio to bypass commercial communication networks is prohibited [4].

Today, amateur radio uses analog transmissions such as Morse code (CW), single sideband (SSB), and frequency modulation (FM), and a wide range of digital modulations such as frequency shift keying (FSK), phase shift keying (PSK), and 8-GFSK (Gaussian FSK) for radio teletype (RTTY), amateur slow scan television (ATV or SSTV) and weak signal Joe Taylor (WSJT) applications such as meteor scatter, moon bounce, propagation sounding, or other types of weak signal work. Morse code is still a very popular communications mode in ham radio, despite the fact that most countries removed the code proficiency requirement for licensing decades ago. A computer communications mode known as FT-8 has become extremely popular in recent years and is used for DX computer communications with very modest stations or when propagation is marginal [5]. Casual over-the-air listening (e.g., tuning of the bands) quickly reveals that rig styles vary widely, from antique tube radios to ultra-modern direct conversion software defined radios (SDR) that hams build from scratch (e.g., home brew), purchase in kit form and assemble, or purchase from a wide range of international vendors. Hams design and deploy their own antennas for their rigs, and operate from a wide range of locations. Antennas run the gamut from simple indoor wire dipoles for apartment dwellers, to whip antennas on vehicles, to massive towers with rotatable yagi beams erected on large seaside lots or mountaintops. Some hams operate large stations remotely by logging in from anywhere via the Internet.

IN THE BEGINNING

At the dawn of wireless, none of today's modulations existed. There was only the spark gap transmitter and the coherer receiver, and only Morse code telegraphy could be used, since Prof.

Reginald Fessenden had not yet stunned and amazed the world with his transmission of sound. In 1899, Prof. Jerome Green of Notre Dame published "The Apparatus for Wireless Telegraph" in the July edition of *American Electrician*, where he provided the reader with very modest but detailed circuitry using commonplace electronics to transmit up to two miles [6]. At that time, radios were constructed on blocks of wood with hand wound coils, crudely built capacitors (condensers), and batteries, as can be seen in Fig. 1 from Marconi's original wireless set from 1894.

Professor Green's paper launched the hobby of ham radio by igniting the imagination of experimenters through the magic of wireless (Fig. 2), and that magic continues to fuel the hobby to this day. As shown in this article, governments had to quickly catch up to encourage but manage the use of this new thing called "radio spectrum," as more and more experimenters began to build transmitters to tinker with wireless communications. As discussed subsequently, the sinking of the Titanic in 1912 brought the entire world together with the realization that it was in the public's best interest to encourage all possible development in wireless communications.

The growing fascination with and interest in amateur radio was not just due to the birth of wireless and the development of affordable electronics alone. The Wright Brothers' first flying machine, demonstrated in 1903, followed by their successful airplane flight of 1905, further fueled intense interest in wireless. Now, hobbyists with an inclination for understanding how things worked (e.g., those interested in engineering) could simultaneously experiment with small model airplanes as well as wireless sets, and many hobbyists realized that wireless could eventually be used to control airplanes and other remote devices.

These early tinkerers, the first amateur radio enthusiasts, operated in an unregulated world without any channelized spectrum, no standards, and no licenses, just wide-open opportunities to experiment, transmit, receive, and tinker with the crude electronics of the day. Anyone was able to communicate with anyone else who happened to also be using the radio waves at the same time, despite the public's perception that wireless telegraph was private [7]. A 1907 article in *Electrician and Mechanic* taught readers how their amateur stations could be expected to perform, based on the particular circuitry, antenna, and detector they chose to use (Fig. 3). See Table 1 From 1907 Electrician and Mechanic (From: [8]).

It was also in 1907 that the U.S. Navy became particularly bothered by two teenagers, Henry C. Heim and Alfred Wolf, who listened in to the wireless telegraph transmissions between a fleet of naval vessels off the shore of Alameda, California. The boys heard messages and compiled a book of the "choicest confidences" between sailors, officers, and people on land, and then provided the most embarrassing messages to the *San Francisco Examiner*. The boys even went so far as to spoof an Admiral, using their clandestine transmitter to delay the departure of a naval ship [9]. These types of incidents gave early amateur operators a bad name, and created deep mistrust among naval and commercial operators while the hobby was in its infancy. However, even with such occurrences, the public's growing interest in radio could not be squelched. The ability to create one's own equipment to navigate the invisible airwaves fueled further interest in this fascinating hobby [10].

It was the dual and near-simultaneous discoveries of practical wireless communications and airplanes that motivated a small group of teenagers from New York City to transform their Aero Club to the Junior Wireless Club in January 1909. This group of teenaged and pre-teen boys, led by honorary president E. Lillian Todd (the world's first woman aircraft designer) and advisor Prof. Reginald Fessenden, eventually became the Radio Club of America, the world's oldest radio society [8]. The title of "first wireless society in the world" is sometimes contested by the Australian Institute of Wireless, founded in 1910 as the Wireless Institute of New South Wales, later becoming the Wireless Institute of Australia. The Wireless Association of America, an honorary club founded in 1909 by Hugo Gernback, who published *Modern Electrics* (the world's first magazine to cover strictly radio topics), was launched just weeks after the Junior Wireless Club, and boasted 10,000 members by late 1910. Local wireless clubs sprang up all over the world in the 1909–1912 timeframe, providing the hobby with enough "grass roots" to influence governments and the public about the coming age of wireless communications and the need to foster the amateur radio hobby.

FIGURE 1. The author with Guglielmo Marconi's original wireless set as preserved from 1894 at the Marconi museum in Villa Griffone, Balogna, Italy (photo taken in October 2013 at the Marconi Society board meeting).

As just one example of the influence of the very earliest of amateur radio enthusiasts, the Junior Wireless Club (later known as the Radio Club of America) was a driving force that caused the U.S. Congress to reject Senator Depew's 1910 proposed legislation that would have completely prohibited amateur radio operators in favor of purely commercial and military (e.g., Navy) spectrum use. The lads went to Washington, DC to testify before a senate subcommittee to plead their case for the importance of amateur radio spectrum [11, 15], and other radio societies lobbied elected officials in force [8]. Students at universities across the world were beginning to build amateur radio sets, while naval officials, upset with the amateurs, complained at congressional hearings that "it was quite common with youths living in the bay cities to play with wireless telegraphy, having aerials and receiving apparatus by which could be read government wireless messages" [9]. It was clear that military and commercial interests were dead set on prohibiting amateur radio operators from ever accessing or listening in to the radio spectrum. Bill after bill in Congress was proposed from 1909 through 1912, only to be rejected by elected officials who were bombarded by the impassioned pleas for spectrum access by members of fledgling wireless clubs across the country.

The sinking of the Titanic on April 14, 1912 was the watershed moment for amateur radio. The Marconi telegraph was credited for saving over 700 people at sea, and all of humanity could now see the importance of wireless and the concurrent need to cultivate expertise in the field of wireless. This event drove Congress to modify their legislative approach. Instead of banning amateur operators altogether, the legislators agreed to finally set aside dedicated radio frequencies for experimentation and hobby use, knowing full well that the spectrum being allocated to the hams was far beyond the realm of usefulness for the technology used at that time. When President William Howard Taft signed the Radio Act of 1912 on August 17, amateur radio operators finally had a law to permit their operation, but they were relegated to the very upper, unexplored reaches of the radio spectrum — the wavelengths of 200 m and smaller —which had yet to be understood or harnessed for use [8, 11]. This model of free, protected radio spectrum for amateur hobbyists was first implemented in the United States, and was eventually adopted worldwide, leading to today's ITU global allocations and protections for the amateur radio service, despite

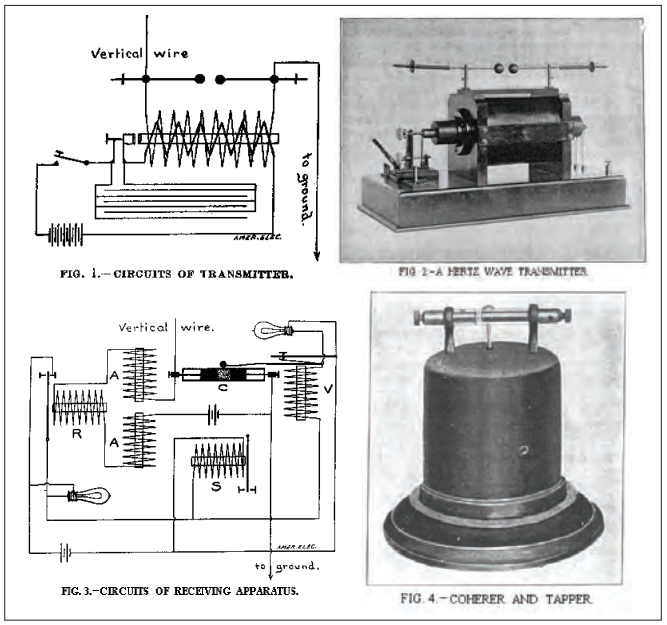


FIGURE 2. Circuitry published by Prof. Jerome Green in the July 1899 issue of *American Electrician* [6]. The circuits were able to cause the coherer to ring a bell using wireless transmissions over a distance of 2 miles on the campus of Notre Dame. https://earlyradiohistory.us/1899nd.htm.

Spark Coil	Antenna Height	Type of Detector: Coherer	Type of Detector: Liquid Detector or Barretter
½ inch	35 feet	⅛ mile	¼ to ½ mile
1	40–45	¼	¼ to ¾
2	50	2½ to 3½	5 to 10
4[1]	75	10	10 to 20
6[1]	100	15	15 to 30
10[1]	150	50	50 to 75
15[1]	180	60 to 75	75 to 100

[1] Tuned.

FIGURE 3. A table from the *Electrician and Mechanic* in 1907 teaches readers what type of range can be expected when constructing their home brew wireless transmitter and receiver system [8].

overwhelming commercial and military interests in radio spectrum.

This type of story, where hams have been the very early adopters — the first tinkerers and the ones to push the envelope of a new technology, often before they become adults — has played out time and time again since the dawn of wireless, and continues today. As shown in this series, from the dawn of the information age, from the early days of talking movies, through the golden age of AM and FM broadcasting, through the advent of television, satellites, cellular telephones, packet radio, and the Internet, hams have always been at the very cutting edge, and have in many cases been responsible for creating the continuing advances in communications. Given the heritage of ham radio operators as early adopters (and in many cases the inventors) of new technologies and communication modes, it should come as no surprise that vendors with pioneering ideas often target the amateur radio market before investing and growing product lines for government and commercial marketplaces. As described in this series, it is the ham radio community's spirit of tinkering with, improving, and socializing each new communications technology that has brought high tech solutions to the world at large.

The Ham Spirit: A Global Fellowship

The ham radio hobby has an incredibly strong "esprit de corps," since licensed amateurs throughout the world realize their good fortune of having free access to the very precious resource of spectrum that is usually licensed to commercial interests to generate revenues for government coffers. From this sharing of the "radio commons," a genuine comradery and deep respect among fellow hams permeate the hobby, transcending language barriers, socioeconomic status, religion, level of education, profession, or country borders. Women and young people (teenagers or younger) are especially encouraged and welcomed both on the air and at local in-person monthly meetings held at the tens of thousands of community amateur radio clubs throughout the world. But the hobby realizes it must do more to bring in underrepresented citizens. In a February 2022 opinion article, David Minter, the Chief Executive of the American Radio Relay League (ARRL), one of the world's largest national amateur radio organizations, made a call for an increased push to increase the diversity of the hobby in the face of an anemic 1 percent annual growth rate for U.S. amateur licensees [12]. Today, about 15 percent of licensed ham operators are female, and many males are aging out of the hobby. With the great need to increase access and diversity in STEM, efforts are underway worldwide to increase underrepresented groups. Within the past few years, the hobby has seen rapid growth of licensed amateur radio operators in Thailand, China, and Indonesia, which bodes well for the international growth of the hobby.

Hams are never strangers to one another — there is most often an instant bond when fellow hams meet for the first time, whether in person or on the air — and life-long friendships across the world naturally evolve through shared interests in the hobby. Ham operators who talk with each other over the radio may exchange electronic or paper QSL cards commemorating on-air contact, although most QSLing today is done through the Internet, using applications such as the Logbook of the World developed by the American Radio Relay League (ARRL).[1]

Hams enjoy meeting over the airwaves, but also often meet in person for technical interchanges or to simply rag chew about their stations, recent contacts, or matters entirely unrelated to ham radio. Such meetings are usually over a meal, and when such a meeting occurs in person, as opposed to over the air, it is called an "eye-ball QSO." It is not uncommon for hams to talk with each other over the air for years, perhaps never meeting in person, or perhaps meeting their invisible ham friend in person years or decades later, at a distant club meeting or "hamfest."[2]

Many life-long, deep personal and professional linkages are formed between fellow hams that typically transcend the hobby, leading to planned family vacations, job opportunities, technical innovations, and business ventures, all while creating a vast citizenry with core competencies and vital expertise that propels the entire electronics communications complex in peace time, or the military complex during war time. It is not an exaggeration to state that ham radio was the world's first global social network.

Some Famous Amateur Radio Operators

The ranks of radio amateurs are filled with women and men who have shaped our civilization from all walks of life. While it is impossible to list the enormous number of hams who have made history, a small sampling will motivate the reader to see how the hobby involves people from many different backgrounds. For example, famed Austrian-born movie actress Hedy Lamarr (Fig. 4) invented frequency hopping spread spectrum and created underwater missile guidance systems while dabbling in amateur radio before World War II [13].

[1] The international amateur radio community uses Q-signals and other abbreviations for common messages. For example. a radio contact is called a QSO, a ham's location is called their QTH, and a station may make an open call to solicit a radio contact on any frequency by calling CQ, (abbreviation stemming from "seek you").

[2] Hamfests are large gatherings of amateur operators and non-licensed radio enthusiasts that are held over a weekend and include technical talks, social outings, swap meets, vendor exhibitions, and flea market equipment sales. The world's largest hamfests often have more than 30,000 attendees and are held annually in Dayton, Ohio in May, Central Florida in February, Friedrichshafen, Germany in June, and Tokyo in October, with thousands of other smaller hamfests held throughout the world each year.

Other notables such as Rajiv Ghandi (Prime Minister of India, VU2RG), King Hussein of Jordan (JY1), King Juan Carlos of Spain (EA0JC), King of Thailand Bhumibol Adulyadej (Rama IX, HS1A), King Hassan II of Morroco (CN8MH), the Sultan of Oman, Qaboos bin Said al Said (A41AA), Carlos Saul Menem (President of Argentina, LU1SM), U.S. Senator Barry Goldwater (K7UGA), movie actor Marlon Brando (FO5GJ), Yuri Gagarin (the Russian astronaut who was the first human in space, UA1LO), Owen Garriott (first ham radio astronaut to transmit from space on the space shuttle, W5LFL), Mamoru Mohri (the first Japanese space program astronaut, 7L2NJY), Helen Sharman (the first British astronaut in space, GB1SS), Kathy Sullivan (the first NASA astronaut to walk in space and to dive to the deepest part of Earth, N5YYV), aviation magnate Howard Hughes (W5CY), Jon Sculley (former president of PepsiCo and CEO of Apple, K2HEP), Alex Comfort (the author of *Joy of Sex*, KA6UXR), and baseball great Joe Rudi (NK7U) are or have been licensed hams. Many famous musicians and recording artists are or were amateur operators, including Jim Croce, Burl Ives, Joe Walsh of the Eagles, teenage idol Donnie Osmond, Patty Loveless, Chet Atkins, Ronnie Milsap, Larnelle Harris, and Larry Junstrom. Even the wife of Elvis Presley, Priscilla Presley, is a ham radio operator, while the legendary country music star Johnny Cash was a highly skilled Morse code interceptor in the U.S. Air Force before he became a country music icon.

Ham radio has influenced all segments of radio and television, providing a crucible from which the personalities and technical innovations for all forms of broadcast entertainment have been hatched. Hams have ignited the world's imagination and fascination with radio since the golden age of radio in the 1920s, from the era of silent movies through the global adoption of AM and FM broadcasting, through the black and white, and then color, television era, and through the coverage of the international space race [14–16]. Walter Cronkite (a U.S. television news anchor known as the most trusted man in America, KB2GSD), space reporter Roy Neal (K6DUE), and TV humorist and producer Jean Shephard (K2ORS) molded the way content was delivered over the commercial airwaves, while Akio Morita (JP1DPJ) and Masru Ibuka (J3BB) founded Sony Corporation, Robert Moog (K2AMH) invented the music synthesizer, Ray Dolby (F5BVY) invented cinematic sound at Dolby Labs, Leo Fender (W6DUE) created the Fender electric guitar, and Nolen Bushnell (W7DUK) founded video game pioneer Atari.

Along with the large number of high-tech entrepreneurs who used amateur radio to kindle their early love for communications and electronics, the famous science fiction writer, Sir Arthur C. Clarke, conceived the notion of geostationary satellites in 1945 [17], and had personal interest and deep connections with the amateur radio community in Sri Lanka. The famous founders of Apple, Steve Jobs and Steve Wozniak, honed their high-tech computer instincts through ham radio as young boys [18, 19], where they greatly benefitted from the Silicon Valley culture inculcated by the venerable Dean of Engineering of Stanford University, Frederick Terman, who founded Silicon Valley three decades earlier. Terman was also a ham radio operator in his youth, much younger than others practicing the hobby in his home town [20], and his early experiences interacting with adults, while he himself was a child, developed both his engineering skills and confidence to later lead large teams to form companies while also creating the classic electrical engineering education textbooks of his era that impacted the entire field of engineering. Terman's students, Bill Hewlett and David Packard (also hams), founded Hewlett Packard in 1939, which became the global leader in test equipment manufacturing. The first Hewlett Packard audio oscillator was used for Walt Disney's groundbreaking movie *Fantasia* — the first movie to be shown with stereo sound — in 1940. Today's corporate giants Hewlett Packard Enterprises, Aruba, Agilent, and Keysight Technologies can all trace their roots to the two amateur radio operators who studied under their Stanford professor who, too, was a ham.

FIGURE 4. Movie star Hedy Lamarr was an avid experimenter with radio and invented frequency hopping spread spectrum (Photo from https://www.forbes.com/sites/shivaunefield/2018/02/28/hedy-lamarr-the-incredible-mind-behind-secure-wi-fi-gps-bluetooth/?sh=4f411b1341b7.)

FIGURE 5. Al Gross, W8PAL, the father of citizens band radio, is shown with his original walkie-talkie invention that now resides in the permanent collection of Virginia Tech [22]. (Photo from: https://lemelson.mit.edu/award-winners/al-gross).

Jack Kilby, also a ham operator who worked on early transistors, is known as the inventor of the microchip, the hand calculator, and the thermal printer while working at Texas Instruments. Kilby received the Nobel Prize in Physics. Before moving to Texas Instruments, however, Kilby worked at Centralab in Milwaukee where he collaborated with another ham operator, walkie-talkie inventor and citizens band pioneer Al Gross, to create a rugged circuit board capable of housing tubes for the world's first walkie-talkies that Gross developed and sold to the public in 1938 [21–23]. More recently, Princeton astrophysicist and Nobel Prize winner Joe Taylor (K1JT) has developed open source weak signal modulation and coding software that has caught the ham radio hobby by storm over the past decade, as it enables ham operators to make QSOs around the world using many exciting propagation channels using very modest powers and antennas with a simple sound card and computer-controlled digital modulations [5].

Conclusion and What's to Come in This Series

This article, Part 1 of a three-part series, has provided an introduction to the origins and current practice of the hobby of amateur radio. It should be clear that the hobby has had a great impact on the communications engineering profession, and offers a special comradery that exposes its practitioners to many social interactions and technical development opportunities with experts and enthusiasts from all walks of life. The ability to expand one's technical knowledge and interest has always been at the core of the hobby. In the coming series of articles, the historic account of how amateur radio operators and the hobby played a key role in the creation of today's information age will be chronicled.

Part 2 of this series will delve into the history of how universities facilitated amateur radio on their campuses to build up a major arsenal of technical experts who went on to develop the global radio broadcast industries, long-distance telephone, television, stereo, the walkie-talkie, and radar. The importance of amateur radio in exploring the Earth by sea, and the role of hams in World War I and World War II will also be described, as militaries relied on the skills and innovations of amateur operators to support wartime communications. A historical perspective of the growth of Silicon Valley, and other high-tech centers around the world will highlight the role that ham radio played, as these technical incubations and innovations were couched in the amateur radio spirit, with hams often being the first employees of startup companies that powered the expansion of the information age. Even during the Cold War, the history will show how amateur radio maintained international relations among ham radio enthusiasts on both sides of the Iron Curtain, even as governments were adversarial to each other.

Part 3, the final installment in this series, will offer a historical account of how amateur radio operators began learning about VHF and UHF frequencies, where they created nationwide repeater systems, often with touch-tone telephone capabilities, thereby proving the concepts and creating the global engineering talent pool needed for the fledgling cellular telephone industry. The history of the Internet would not be complete without understanding how amateur operators developed and perfected packet data communications, digital modulations, and open source circuit boards and software for nationwide computer networks that linked thousands of amateur stations across the world, decades before the Internet. The activities of pioneering ham clubs like Tucson Amateur Packet Radio (TAPR) sparked global interest in computer networking through the ham radio "packet cluster." The "big board" computers built at ham clubs across the world in the 1970s and 1980s fueled the fledgling personal computer industry. The launch of OSCAR 1 in 1961, the world's first amateur radio satellite, spawned deep technical expertise and technical know-how for the global satellite industry, leading to the creation of non-profits such as Radio Amateur Satellite Corporation (AMSAT), which sponsored rockets and built satellites that serve hams around the world. Hams of every nation participate in their country's "field day," where clubs set up stations in the wilderness over a weekend to simulate emergency communications needed in a national disaster [24]. The hobby even has its own Olympic Games. Known as the World Radiosport Team Competition (WRTC), the ham radio Olympics are held every four years around the world, and will next be hosted in 2023 by Italy in the city of Bologna (the site of Marconi's original wireless transmissions) [25]. The use of amateur radio to explore the ionosphere in new ways, and with new radio architectures and crowd-sourcing Internet-based monitoring systems like the Reverse Beacon Network (RBN), has created new fields of research, such as "space weather" to predict the upper reaches of our planet's protective layers. The final article in this trilogy shall delve into the fascinating role that ham radio has played in all of these modern advances that impact civilization, and shall conclude with a perspective on what lies ahead for the incredible hobby of amateur radio.

References

[1] D. Haldane, "Girl Hams It Up for the World : Ham Radio: At 5, She's Maybe the Youngest Operator in U.S. Her Mental Skills Are Surprising," *Los Angeles Times*, July 21, 1991; https://www.latimes.com/archives/la-xpm-1991-07-21-hl-502-story.html.

[2] "Oldest Known US Radio Amateur, Cliff Kayhart, W4KKP, SK at 109," *QST Magzine, American Radio Relay League News*, Oct. 29, 2020; http://www.arrl.org/news/oldest-known-us-radio-amateur-cliff-kayhart-w4kkp-sk-at-109

[3] "Amateur and Amateur-Satellite Services," Radio Regulations, ITU-R, *Radio Handbook*, 2014; https://www.itu.int/en/publications/ITU-/pages/publications.aspx?parent=R-HDB-52-2014&media=electronic).

[4] J. Pepitone, "Is Ham Radio A Hobby, A Utility... or Both?," *IEEE Spectrum*, July 8, 2019; https://spectrum.ieee.org/is-ham-radio-a-hobby-a-utilityor-both-a-battle-over-spectrum-heats-up.

[5] J. Taylor, K1JT, "WSJT Home Page by K1JT"; https://physics.princeton.edu/pulsar/k1jt/, accessed Aug. 1, 2022.

[6] J. Green, "The Apparatus for Wireless Telegraph," *American Electrician*, July 1899.

[7] A. C. Madrigal, "The Great Wireless Hack of 1903," *The Atlantic*, Dec. 9, 2011; https://www.theatlantic.com/technology/archive/2011/12/the-great-wireless-hack-of-1903/250665/.

[8] C. B. DeSoto, "Two Hundred Meters and Down: The Story of Amateur Radio," The American Radio Relay League, 1936, West Hartford, CT.

[9] Hearings Before a Congressional Subcommittee of the Committee on Naval Affairs of the House of Representatives on H.J. Resolution 95, A Bill to Regulate and Control the Use of Wireless Telegraphy and Wireless Telephony, Government Printing Office, 1910, Washington, DC.

[10] R. A. Bartlett, *The World of Ham Radio, 1901–1950: A Social History*, MacFarland and Co., 2007.

[11] W.E.D. Stokes, Jr. "The Radio Club of America," *Proc. Radio Club of America (1BCG Commemorative Issue)*, Oct. 1950, pp. 66–67; https://worldradiohistory.com/Archive-Radio-Club-of-America/Radio-Club-of-America-1950-10.pdf.

[12] D. A. Minster, NA2AA, "Diversity and Inclusion: Driving Amateur Radio's Growth," *QST Mag.*, Feb. 2022, p. 9.

[13] D. Khan, "Cryptology and the Origins of Spread Spectrum: Engineers During World War II Developed an Unbreakable Scrambler to Guarantee Secure Communications Between Allied Leaders; Actress Hedy Lamarr Played A Role in the Technology," *IEEE Spectrum*, vol. 21, no. 9, pp. 70–80; https://ieeexplore.ieee.org/document/6370466.

[14] S. J. Douglas, *Listening In: Radio and the American Imagination*, Times Books, 1999.

[15] *Proceedings of the Radio Club of America*, Diamond Jubilee issue, 75th Anniversary of the Radio Club of America, 1984.

[16] T. Lewis, *Empire of the Air: The Men Who Made Radio*, Harper Collins, 1991.

[17] A. C. Clarke, "V2 for Ionospheric Research?," Letters to the Editor, *Wireless World*, Feb. 1945, p. 58.

[18] Razvan, "Steve Jobs Influenced by Ham Radio, "Nov. 27, 2013, based on Smithsonian Institute/Computerworld Information Technology Awards History Project, April 20, 1995; https://qrpblog.com/2013/11/steve-jobs-influenced-by-ham-radio/ (also find the Oral Histories at Smithsonian shown on this webpage for better cite).

[19] W. Isaacson, *Steve Jobs*, Simon & Schuster, 2012.

[20] San Jose History Project, Frederick Terman with amateur radio equipment, 1917; https://historysanjose.pastperfectonline.com/photo/AF0C3384-3C14-44E2-BFED-154452200970; also see https://calisphere.org/item/208eedb7acabd2b9f3e6f85b1ebd0729/

[21], "In Memorium: Alfred J. Gross, Wireless Radio Systems Pioneer," *IEEE AESS Systems Mag.*, Mar. 2001, pp. 43–44; https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=911321.

[22] A Guide to the Al Gross papers, 1909–2000, Virginia Tech Library, A Collection in Special Collections, Collection Number Ms2001-011; https://ead.lib.virginia.edu/vivaxtf/view?docId=vt/viblbv00043.xml.

[23] Correspondence with Al Gross, from Jack S. Kilby Papers, 1878–2003, U.S. Library of Congress; https://www.loc.gov/item/mm2006085317/.

[24], "The Uncertain Future of Amateur Radio," *IEEE Spectrum*, July 10, 2020; https://spectrum.ieee.org/ham-radio.

[25] J. K. George, "Contact Sport: A Story of Champions, Airwaves, and a One-Day Race around the World, " c. 2016.

Biography

Theodore. S. Rappaport (tsr@nyu.edu) is the David Lee/Ernst Weber Professor at New York University (NYU). He founded the NYU WIRELESS research center and the wireless research centers at the University of Texas Austin (WNCG) and Virginia Tech (MPRG). His work has provided fundamental knowledge of wireless channels used to create the IEEE 802.11 standard, the first U.S. digital TDMA and CDMA standards, the first public Wi-Fi hotspots, and recently proved the viability of mmwave and sub-THz frequencies for 5G, 6G, and beyond. He founded two businesses that were sold to publicly traded companies – TSR Technologies, Inc. and Wireless Valley Communications, Inc., and was an advisor to Straight Path Communications which sold 5G mmWave spectrum to Verizon. He is a licensed Professional Engineer and is in the Wireless Hall of Fame, a member of the National Academy of Engineering, a Fellow of the National Academy of Inventors, and a life member of the American Radio Relay League. His ham radio call sign is N9NB.

CALL FOR PAPERS

IEEE COMMUNICATIONS MAGAZINE

FEATURE TOPIC: METAVERSE

BACKGROUND

The metaverse is an emerging concept which embraces three important elements. First, it embraces a social element, i.e. it is not only a virtual space where users spend time (and money) on their own or with a selected few. Rather, the metaverse is intended to resonate with the very social fabric which underpins human society. Once in the metaverse, you and/or your avatar are able to interact humanly by looking into each other's eyes, perceive body language and maybe even shake hands.

Second, it has a strong virtual narrative where our experiences and ways of social interaction are significantly augmented with persistent virtual content, consumed via VR or AR. Access to the virtual world of the metaverse and haptic interaction therein is enabled by eXtended Reality (XR) devices, and in the interim via today's screens leveraging WebXR technologies.

Third, it is accelerated through novel technologies, like Web 3.0 and blockchains, digital twins, real-world graphics rendering, artificial intelligence and 5G/6G, just to name a few.

After decades of technology enabling personalization, perhaps now – with the emergence of the metaverse – we can build a native technology that allows us to interact and share realities together with our surroundings. Augmenting our reality while blending it with the digital world underpins a shared reality where people come together instead of apart. Digital sharing via socials caused us sharing less in our real lives. Immersive tech can be the inflection point where ;we share and socialize more in our real life in a healthier and more sustainable way. We also know that the paradigm shift from "content inside a screen" to "users inside the content" needs to be inclusive of people who can't do wearables or who can't do embodied interactions. This calls for new ways of enabling content flow, infrastructure, accessibility that will ultimately redefine our ecosystems and daily activities, work and social lives.

This Feature Topic (FT) aims to bring together researchers, industry practitioners, and individuals working on the related areas to share their new ideas, latest findings, and state-of-the-art results. Prospective authors are invited to submit articles on topics including, but not limited to:

- Metaverse services and business models
- Metaverse architectures
- 6G techniques, algorithms and architectures
- User experience, avatar
- AR/VR devices
- XR haptic devices
- Metaverse standards
- Metaverse media codecs
- Web 3.0, blockchain
- Ethics, Privacy, Security
- Sustainability, Social Good
- Experimental demonstrations and prototypes

SUBMISSION GUIDELINES

Manuscripts should conform to the standard format as indicated in the Information for Authors section of the *IEEE Communications Magazine's* Manuscript Submission Guidelines. Please, check these guidelines carefully before submitting since submissions not complying with them will be administratively rejected without review.

All manuscripts to be considered for publication must be submitted by the deadline through Manuscript Central. Select the "FT-2219/Metaverse" topic from the drop-down menu of Topic/Series titles. Please observe the dates specified here below noting that there will be no extension of submission deadline.

IMPORTANT DATES

Manuscript Submission Deadline: 31 October 2022
Decision Notification: 31 January 2023
Final Manuscript Due Date: 28 February 2023
Publication Date: First Quarter/Second Quarter 2023

GUEST EDITORS

Mischa Dohler
Ericsson Inc, Silicon Valley, USA
mischa.dohler@ericsson.com

Mar Gonzalez Franco
Microsoft Research, Redmond, USA
margon@microsoft.com

Young-Gab Kim
Sejong University, Korea
alwaysgabi@sejong.ac.kr

Timoni West
Unity AR/VR Chief, USA

A Look Back 10 and 25 Volumes Ago

IEEE Communications Magazine
October 1997, Volume 35, No 10

Management of Broadband Networks

Intranet Services and Communication Management

Topics in Personal Communications

IEEE Communications Magazine
October 2012, Volume 50, No 10

Topics in Military Communications

Topics in Automotive Networking and Applications

Topics in Integrated Circuits for Communications

Accepted from Open Call

UPDATED ON THE COMMUNICATIONS SOCIETY'S WEB SITE

www.comsoc.org/conferences-events/portfolio-conferences-events

2022

OCTOBER

IEEE CTW 2022 — IEEE Communications Theory Workshop 2022, 2–5 Oct.
Marbella, Spain
https://ctw2022.ieee-ctw.org/

NoF 2022 — 13th Int'l. Conference on Network of the Future 2022, 5–7 Oct.
Ghent, Belgium
https://nof2022.dnac.org/

WCSP 2022 — Int'l. Conference on Wireless Communications and Signal Processing 2022, 14–17 Oct.
Nanjing, China
http://www.ic-wcsp.org/2022/

CCCI 2022 — Int'l. Conference on Communications, Computing, Cybersecurity, and Informatics 2022, 17–19 Oct.
Virtual
http://atc.udg.edu/CCCI2022/

IEEE HEALTHCOM 2022 — IEEE Int'l. Conference on E-health Networking, Application & Services 2022, 17–19 Oct.
Genoa, Italy
https://healthcom2022.ieee-healthcom.org/

ICTC 2022 — Int'l. Conference on Information and Communication Technology Convergence 2022, 19–21 October 2022
Jeju-si, Jeju-do, South Korea
http://ictc.org/

APCC 2022 — Asia Pacific Conference on Communications 2022, 19–21 October 2022
Jeju-si, Jeju-do, South Korea
http://apcc2022.org/

IEEE SmartGridComm 2022 — IEEE Int'l. Conference on Communications, Control, and Computing Technologies for Smart Grids 2022, 25–28 October 202
Singapore/Hybrid
https://sgc2022.ieee-smartgridcomm.org/

WINCOM 2022 — 9th Int'l. Conference on Wireless Networks and Mobile Communications 2022. 26–29 Oct.
Rabat, Morocco
https://www.wincom-conf.org/WINCOM_2022/

NOVEMBER

IEEE CAMAD 2022 — IEEE Int'l. Workshop on Computer Aided Modeling and Design of Communication Links and Networks 2022, 2–3 Nov.
Paris, France
https://camad2022.ieee-camad.org/

IEEE CloudNet 2022 — IEEE Int'l. Conference on Cloud Networking 2022, 7–10 Nov.
Paris, France
https://cloudnet2022.ieee-cloudnet.org/

IEEE NFV-SDN 2022 — IEEE Conference on Network Function Virtualization and Software Defined Networks 2022, 14–16 Nov.
Chandler, AZ
https://nfvsdn2022.ieee-nfvsdn.org/

ITNAC 2022 — Int'l. Telecommunication Networks and Applications Conference 2022, 23–25 Nov.
Wellington, New Zealand
https://itnac.org.au/

MILCOM 2022 — Military Communications Conference 2022, 28 Nov.–2 Dec.
Rockville, MD
https://milcom2022.milcom.org/

IEEE LatinCom 2022 — IEEE Latin-American Conference on Communications 2022, 30 Nov.–2 Dec.
Rio de Janeiro, Brazil
https://latincom2022.ieee-latincom.org/

DECEMBER

IEEE GLOBECOM 2022 — IEEE Global Communications Conference 2022, 4–8 Dec.
Rio de Janeiro, Brazil/Hybrid
https://globecom2022.ieee-globecom.org/

Communications Society portfolio events appear in bold colored print. ▪ Communications Society technically co-sponsored conferences appear in black italic print.

GCN GLOBAL COMMUNICATIONS NEWSLETTER

October 2022
ISSN 2374-1082

INTERVIEW

Asia Pacific Region

Interview with Tomoaki Otsuki, Director of the AP Region

by Stefano Bregni, Global Communications Newsletter Editor-in-Chief, Director Conference Operations, and Tomoaki Otsuki, Director of the AP Region

This article continues the series of nine interviews to the Officers of the IEEE ComSoc Member and Global Activities (MGA) Council, which is published every month on the Global Communications Newsletter.

In this series of articles, I introduce the Vice-President and the six Directors on the MGA Council (namely: Member Services, Industry Outreach, and AP, NA, LA, EMEA Regions), as well as the two Chairs of the Women in Communications Engineering (WICE) and Young Professionals (YP) Standing Committees. In each interview, one by one they present their sector activities and plans.

In this issue, I interview Tomoaki Otsuki, IEEE ComSoc Asia Pacific Region Director.

Stefano Bregni

Tomoaki Otsuki

Tomoaki is a professor at the Keio University in Yokohama, Japan. He is the author or co-author of over 235 research papers and 470 international conference papers, in the areas of wireless communications, optical communications, signal processing, sensing, and so on. He is now serving as an Area Editor of the *IEEE Transactions on Vehicular Technology* and an editor of the *IEEE Communications Surveys and Tutorials*. He is a senior member and a distinguished lecturer of the IEEE, a fellow of the IEICE, and a member of the Engineering Academy of Japan.

Tomoaki, we might begin outlining the main characteristics of the AP Region.

The AP Region covers a vast geographical area stretching from South Korea and Japan in the north — east to New Zealand in the south, and Pakistan in the west. With a membership of 107,154, it is one of the largest regions in IEEE. This is roughly 40% of ComSoc members in the world.

Chapters provide a local connection for our society members. Their activities include: talks organized within the Distinguished Lecturer Tour (DLT) or Distinguished Speaker Program (DSP) frameworks, social events, workshops, seminars, special events, etc.

The Distinguished Lecturer Program (DLP) and the Distinguished Speaker Program (DSP) are particularly appreciated by our Members. The AP Region is extremely active also on this program.

DLTs provide the means for ComSoc chapters to identify and arrange lectures by renowned experts on communications and networking-related topics. ComSoc's DSPs enable current and past distinguished lecturers as well as ComSoc officers, IEEE Fellows, and prominent speakers to schedule lectures while traveling on business trips. To cope with the vast geographical area, and related travel costs, of the AP Region, we are emphasizing DLT/DSP.

What about other membership activities in the AP Region?

We also put lots of efforts to promote our membership activities, particularly women in communication engineering (WICE), young professionals (YP), students, and industry activities. We have set up a new team including new WICE and YP committees. Dr. Kaoru Ohta of Muroran Institute of Technology, Japan and Dr. Jemin Lee of Sungkyunkwan University (SKKU), Korea are leading the WICE committee. Dr. Nicholas Wong of GlobalFoundries, Singapore, and Dr. Yuichi Kawamoto of Tohoku University, Japan are leading the YP committee. They are planning several events and services for WICE and YP members.

Regarding the industry activity, we organized several DLT/DSP by people from industry. Also, we sponsored the 2nd International Future Communications Workshop (2 day online industry-academia joint workshop) that all our member can attend free.

An important facet of the AP activities is our Award Program that includes the AP Young Researcher Award and the AP Outstanding Paper Award. The former one is 17th and the latter one is 11th this year. Many recipients of the AP Young Research Award including me are now contributing a lot to our IEEE the AP Regional Activities.

The AP Region publishes a very detailed Newsletter. May you please tell us a bit about that?

Another important service is our AP Newsletter. We publish 2 AP newsletters a year. We have already published 61 AP newsletters. The AP newsletter contains a lot of useful information for our members including our award programs and other activities, also sometimes tutorial articles.

And what about Students? Do you organize regularly any specific activity for them?

In AP regions, we are also proud that many chapters in our AP regions are very active. They held many events including many events for students. Thanks to their activities, many chapters increased members a lot, including student members. It is important to share the knowhow with other chapters. We hold the event that introduced the knowhow of the chapter that was very active and the chapter award to other chapter members. We are planning to continue this activity.

How do you coordinate the activities of Chapters? Do you meet regularly with Chapter Chairs?

In the AP region we usually had the AP Region Regional Chapter Chairs Congress (AP-RCCC) every 2 years before the COVID-19. The AP-RCCC is the place where all the chapter chairs get together and discuss our activity. It is a very important opportunity to get the opinion and comments from the chairs directly, which contributed a lot to improve our activities. Unfortunately, we have not held AP-RCCC during the COVID-19, but now we are planning to held it next year in the AP region, possibly at IEEE Globecom 2023 in Kuala Lumpur, Malaysia.

We hope many people will be able to come to the IEEE Globecom 2023!

I am one of the organizers of IEEE GLOBECOM 2023 in Kuala Lumpur and therefore I am extremely happy you have mentioned it. As one of the Technical Program Co-Chairs, I

(continued on next page)

wish to take this opportunity to invite all readers to submit a paper to IEEE GLOBECOM 2023 and, certainly, to attend it in person. KL is an outstanding location.

Changing topic, how the AP Region Board collaborates with other professional associations in the Region?

Another important thing in the AP region is that we have many ComSoc Sister Societies. We held many activities including international conferences and workshops with collaboration with the sister societies.

To promote our activities in local countries, it is very important to have such collaboration with sister societies. We are planning to enhance the collaboration with the sister societies so that we can reach local people more.

To conclude, what are the most serious challenges that you are coping with?

One of our challenges is membership retention. As I mentioned, we have many members in total. However, we have a few Chapters where a significant number of members did not renew the IEEE membership or ComSoc membership. Retention is important for us! We are now investigating the reasons to improve our services to members. We hope we can share our findings with other regions.

INTERVIEW

Communications History: Capturing, Curating and Making Easily Available

Interview with Douglas Zuckerman, IEEE ComSoc Communications History Committee Chair

by Stefano Bregni, Global Communications Newsletter Editor-in-Chief, Director Conference Operations, and Douglas Zuckerman, IEEE ComSoc Past President, Communications History Committee Chair

Today, I have the great pleasure to interview Douglas Zuckerman, IEEE ComSoc Communications History Committee Chair. Doug is not only a great friend of mine, but he also helped me a lot during my first days in the ComSoc Board of Governors, when he was ComSoc President and appointed me Director of Education in the already far 2008.

Dr. Doug Zuckerman, an IEEE Life Fellow, is on the IEEE Future Directions and Industry Engagement Committees. He was IEEE Communications Society President and IEEE Board Director. His BS, MS and PhD degrees are from Columbia University (USA). His earlier work at Bell Labs (and successors) heavily influenced network management standards and implementation.

Stefano Bregni

Douglas Zuckerman

Stefano: Hello Doug! What a pleasure to interview you today. Would you like to begin by introducing what is the Communications History Committee?

Doug: This year marks the 70th anniversary of the IEEE Communications Society, which is its platinum jubilee celebration. This is a singular bit of history, yet could we imagine if we had lost track? What if all the major milestones in communications and in ComSoc had also been either lost or forgotten? Capturing, curating and making easily available this rich history is essential to building and maintaining our professional community. It also helps set a foundation for the future. Thankfully, ComSoc has a standing committee on Communications History, which I currently have the privilege of chairing.

From the Society's Bylaws, "This Committee is responsible for identifying, placing in electronic archives, and raising public awareness through all appropriate means of the most important facts, people, and achievements of communications history, as well as telecommunications milestones in general."

Examples of its activities include solicitation of articles on Communication History for our magazines, organizing communications history sessions at conferences, and commemorating significant anniversaries of the IEEE Communications Society. In 2012, the 60th anniversary of our Society, the Committee organized the updating, expansion and re-issue of the History of Communications book published for our 50th anniversary in 2002, and the production of a 25-minute recollections video in which many former Presidents of IEEE ComSoc recounted their experiences in the communications field and in IEEE ComSoc.

Stefano: What is the current composition of the Communication History Committee?

Doug: Due to the wide range of activities envisioned for this committee, the committee has about a dozen members with the following roster:

Committee Roster

Chair | Douglas Zuckerman (2022-2024)

Voting Members

2022-2024 | Douglas Zuckerman

2022-2023 | Stephen Weinstein

2022 | Scott Atkinson, Kit August, Celia Desmond, Tarek El-Bawab, Rose Hu, Curtis Siller, Martha Steenstrup, Tim Weil, Sarah Kate Wilson

Staff | Cynthis Sikora

This team has former ComSoc presidents, past Communications Magazine EiC's, a past History Committee chair, several Life Members, and founders of some of the earliest ComSoc flagship conferences as well as organizers of more recent ones.

Stefano: Certainly, your Committee includes very well known volunteers of ComSoc, with a long record of service in our Society. What initiatives are you planning for this year?

Doug: Earlier this year, the History Committee came up with the following initiatives for 2022:

- **Communications History Book** — Add events since 60th anniversary edition (also feeds into Interactive Communications History).
- **Interactive Communications History** — Contribute to this project which has been funded by ComSoc's Board of Governors.
- **IEEE Milestones** — Identify and guide proposals for new communications milestones.
- **Conference History** — Start with GLOBECOM/ICC history — including interviews.
- **Presence at IEEE GLOBECOM 2022** — Organize and moderate a panel (currently planned to be a "Panel of Presidents," and will memorialize several leaders who have passed during the year). Have exhibit hall presence.
- **Presence in** *IEEE Communications Magazine* — Do column, reprints, old ComMag synopses, Global Communications News (GCN) articles.
- **Update Web Presence** — Participate in ComSoc-wide website update activity relevant to history activities.
- **Virtual Museum** — Provide an AR/VR museum experience.
- **Diversity, Equity and Inclusion** — Recognize "Women in Communications History."

In addition, there are two other important activities:

- **ComSoc Platinum Jubilee Celebration** — Work with the Board of Governors and ComSoc staff in preparing for the Society's 70th Anniversary Platinum Jubilee celebration.
- **Memorializations** — Write obituaries for ComSoc's "In Memory" website and where appropriate author Society News articles in Communications Magazine for recently deceased colleagues. This year, we have memorialized Bob Lucky and Des Taylor (May they rest in peace.).

(continued on next page)

Stefano: Would you please explain the meaning of the three items you mentioned in the title of this interview? Capturing, Curating and Making Easily Available?

Doug: Capturing, curating and making easily available our history is a daunting, almost boundless, challenge. IEEE itself has an IEEE History Committee with full time staff and has been supporting some of our activities, including hosting an important communications history site: https://en.wikipedia.org/wiki/IEEE_Communications_Society. Discussions are ongoing on how our Society's and IEEE's history committees may continue collaborating on mutually beneficial activities.

An area where we can especially use help is in generating history articles for Communications Magazine. As ComSoc's History Committee chair, I also do a history column in that magazine. I have solicited a few articles that are still in preparation but would like to see a more regular presence. One idea has been to have monthly one-page synopses of the tables of contents from "this month, 25 and 50 years ago." Also, the ComMag EiC has solicited best papers for reprint (with possible historic updates) from issues during their term as EiC. Going forward, I would expect to see more history content appearing in ComMag.

Stefano: Are you working also with some Chapters?

Doug: Since this is an article in the Global Communications Newsletter, I would be remiss if I didn't mention that history also has a geographic perspective. Many IEEE sections and regions have history committees, and these typically include content of relevance to communications. An example that comes to mind is that of the NJ Coast Section. This is the section that had been the host to important fundamental research and development entities in the communications field, e.g., Bell Labs and Fort Monmouth. AT&T also hosts a museum within this section. This example of a section history is at https://ethw.org/IEEE_New_Jersey_Coast_Section_History. Note, its historian is also an active member of the ComSoc History Committee.

Stefano: Besides posting a column on the Communications Magazine, are you also planning to organize some event at ComSoc conferences?

Doug: As already mentioned, we are organizing a "Panel of Presidents" for GLOBECOM 2-22 in Rio de Janeiro this December, as well as being part of the ComSoc exhibition booth. The History Committee is also participating with ComSoc's Board of Governors in identifying other anniversary activities as well.

Above, the Conference History initiative mentioned starting with GLOBECOM and ICC (flagships) capturing their history. This is well underway for our two flagship events. I would like to encourage our important portfolio conferences to also capture their history. Memories fade, so now, at the Society's 70th anniversary, is the time to do it.

For example, how many people will remember when the first NOMS took place, where it was, who was on the organizing committee, and what topics were important at the time? I was able to find the call for participation that appeared in the November 1987 issue of *IEEE Communications Magazine*, when Fred Andrews was ComSoc President (and a strong supporter of NOMS).

The first NOMS took place in February 1988 in New Orleans, was heavily supported by AT&T Bell Labs (especially Eric Sumner) and BellCore (especially Fred Andrews), had the theme, "Productivity Through Operations," was standing room only (registrations were screened and many were turned away), and generated an unmentionably huge surplus. Also, its conference record was "visuals plus accompanying text" (done by "cut and paste" on paper during this pre-PPT era). The attendees were almost entirely from industry. Bruce Kieburtz was General Chair, Ed Glenner was Vice General Chair, and Ken Lutz and I were the Technical Program Committee Chairs. From the call, it is interesting to note the variety of topics that were important to the network operations and management community — and industry — at that time.

Stefano: I am sure you have a number of nice memories to share! Please, choose one from your early career in the communication industry and tell us. Young readers will be interested to hear.

Doug: I could also share early memories from my professional career at Bell Labs, starting at Holmdel on the WT4 Millimeter Waveguide Transmission Systemin June 1969. My first boss told stories of his days on the Telstar project, which was a "spare no money" effort in response to the launch of Sputnik. The WT4 system had promise until long distance fiber proved practical — and then many of us moved onto satellites and undersea cable.

On a more personal level, which will give away my generation, I grew up with vacuum tubes, rotary telephones (on a party line), and our first TV (black and white, with a mechanical tuner, only one for the family). During my undergraduate years at Columbia University (USA), we used a book by Millman and Halkias and taught at the cusp of moving from vacuum tubes to transistors and integrated circuits. I also recall a one-semester class taught by Omar Wing on the Fortran IV programming language - used with punch cards and a huge IBM 360 computer center ("printouts" were overnight at the "economy" rate). Prof. Wing was responsible for my joining IEEE as a student member. In his introductory circuits class, he handed out application forms and said IEEE was good to join. I joined.

This reminds me of an "historic" telecom concept, "The Network is the Database." Analogously, one might say, "The Members are the History." All of us are creating history and storing it in our personal memories (some of which I just shared with you). The challenge is in capturing, curating and making that history easily available. The ComSoc History Committee stands ready to help!

CHAPTER REPORT

News from the Denver ComSoc Chapter

by Tim Weil, ICC 2024 General Chair, and Mark Milliman, Denver ComSoc Chapter Vice-Chair, USA

The Denver COMSOC chapter, like other chapters, is transitioning back to in-person meetings while incorporating some of the positive aspects of the last couple of years. We are facilitating more hybrid meetings especially technical talks to accommodate the hectic schedules of our members. Our section is geographically spread out over a geographical territory that can take two hours to reach from end-to-end. Hybrid meetings provide greater value to our members when the location is far away from them.

Another thing we are doing is co-hosting more events with other societies and geographical chapters. Co-hosting allows us to offer a greater quantity and diversity of technical talks than what we would do alone. Several of our events are applicable to other societies as well. Finally, we are reinvigorating our student chapters. We have three major universities in our section

University of Denver COMSOC Meeting (students and ICC 2024 Committee)
Photo credit - David Gao, Ritchie School of Engineering, University of Denver

that are full of opportunities to connect students with our many local companies.

Our June chapter meeting was sponsored by the University of Denver (DU) as a way to create that opportunity. It was well-attended by DU faculty and students, COMSOC chapter members and the planning committee for the ICC 2024 conference. Part of the program was dedicated to review the recent IEEE International Conference on Communications (ICC 2022 — Seoul, KR) presented by Tim Weil, Denver Chapter officer and chair of the ICC 2024 (Denver) conference.

As General Chair for the ICC 2024 (Denver) program, my recent trip to ICC 2022 seemed like a great opportunity to build interest in our program and recruit new volunteers to our committee. The talk highlighted this year's conference theme "Intelligent Connectivity for Smart World" presenting advances in mobile wireless networks 5G/6G networks, AI/ML research and the future state of high-speed networks.

The meeting was intended to give local chapter members that could not attend or yet to volunteer for planning a better understanding of how the ICC conference is planned, and how they can contribute to our 2024 program. Steve Jia, CableLabs, provided a technical presentation on Radio Access Networks for 5G.

The Denver COMSOC chapter has several social and technical events planned for the rest of the year including company tours, DLT talks, and member presentations as well as a few social events with students and other local society chapters.

The IEEE Denver Section (and COMSOC Chapter) is hosting the 2024 ICC conference in Denver with the theme promoting "Scaling the Peaks of Global Communication Networks." Colorado was home for the early COMSOC flagship conferences including GLOBECOM (1965) and ICC (1993) that were hosted in Boulder and Denver. Our local organizing committee continues to welcome volunteers and local telecommunication industries and professionals to join our program. For more information contact us at — Denver_Chapter@comsoc.org.

Tim Weil at ICC 2022 (Seoul, KR) - https://stories-and-songs.us/node/162

CHAPTER REPORT

Talks on Day of Light at Thailand Chapter

Twenty Five Years Thai Hologram, Solar Farm, and Updated Misleading LiFi

by Keattisak Sripimanwat, IEEE ComSoc Thailand Chapter Chair

As previous years, Thai ComSoc chapter joined the international day of light (IDL) event as a national contact node. For IDL2021, its theme was set with "Society, Energy, and Lighting" (www.quantum-thai.org/idl-2021-thailand). The opening speech was given online on May 16 by Boonrucksar Soonthornthum, an advisor and former director of the national astronomical research institute of Thailand (NARIT). Recent NARIT projects was also introduced.

Following by a full story in order to celebrate the 25th anniversary for the golden jubilee hologram (1996), the documentary behind those beautiful 3D pictures was released by Thai ComSoc chapter. That fascinated collection is not only with historical content, but also includes different R&D perspective with high hidden cost. It is eventually one of the national milestones based on light (IEEE Foundation's granted project 2019). This hologram story is quite a good lesson learned for the future national policy on science and technology development.

Next, the most popular title on "Solar cell & Solar Farm 2021," was given by Dr.Kobsak Sriprapha, a NSTDA's researcher. General knowledge was delivered successfully to non technical background audiences with up-to-date progress of various national projects.

Furthermore, another highlight topic was followed; "Light Fidelity (LiFi) case study: an immunity to pseudoscience." Since LiFi is claimed as a counterpart of WiFi that using LED light bulb to be a new faster hotspot connecting to mobile devices ! However, LiFi as quite a serious hype, has been promoting irresponsibly for over a decade around the world. Thus, revealing its behind stories is that the way to protect our students, members, and people from this kind of crazed engineering. Summary of previous articles (bit.ly/31z0GmT & bit.ly/34Ea36L) with updated misleading stories including the rhetoric & excuses from recent few years, were then presented by the chapter chair. Related science communications through ComSoc's social medias was done throughout the year in order to raising public awareness.

Finally, Thai ComSoc chapter would like to thank all IDL2021's supporters; 1) IEEE photonics society (educational seed funding) 2) institute for the promotion of teaching science and technology (IPST) 3) electricity generating authority of Thailand (EGAT) and 4) metropolitan electricity authority (MEA), for their kindness to co-enlighten people on technology development with fraud prevention.

Poster and video thumbnail of an IDL2021 talk — 25th anniversary of the golden jubilee hologram (1996–2021)

DISTINGUISHED LECTURER

IEEE ComSoc Karachi Chapter

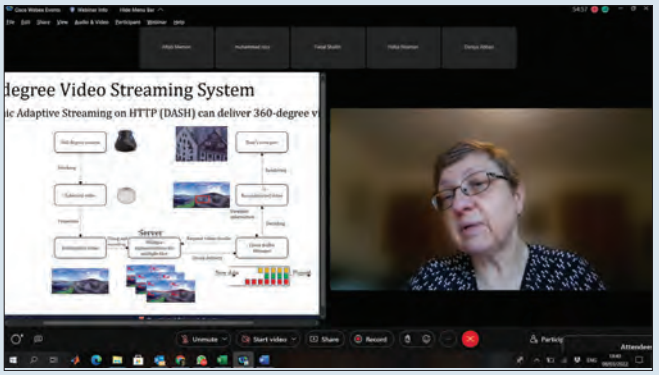
Distinguished Lecture Series on "Navigation Graphs for 360 Degrees Tile Based Video Streaming Systems"

by Umair Ahmed Korai, MUET, IEEE Communication Society Karachi Section Chapter Vice Chair, Pakistan, Bhawani Shankar Chowdhry, MUET, IEEE Karachi Section Chair, Aftab Ahmed Memon, MUET, Pakistan

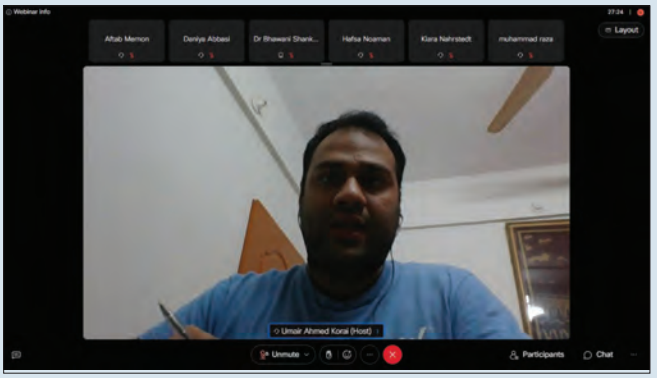
IEEE Communication Society (ComSoc) Karachi Chapter in collaboration with IEEE ComSoc Mehran University of Engineering and Technology (MUET) Student Branch Chapter, IEEE Karachi Section, and QS World Merit organized a second talk of the distinguish lecture series — IEEE ComSoc Karachi chapter on "Navigation Graphs for 360 Degrees Tile Based Video Streaming Systems". The guest speaker of this talk was Prof. Klara Nahrstedt, University of Illinois at Urbana Champaign, United States of America on March 08, 2022. The talk was organized virtually Cisco WebEx tool.

Distinguish Lecture Series — IEEE ComSoc Karachi Chapter is an initiative of IEEE ComSoc Karachi Chapter. The aim of this series is to promote the latest technological information among the students, young professionals, and professionals. The total registered participants in this online talk were 120 out of which 36 were IEEE members and 84 were non-IEEE members.

The online talk was formally started with recitation of the Holy Quran. Afterwards, Dr. Umair Ahmed Korai, Assistant Professor at Dept. of Telecommunication Engineering, MUET and Vice Chair IEEE ComSoc Karachi Chapter, welcomed the worthy Guest Speaker, Chief Guest, Guest of Honor, the Coordinators, the Organizers, and the Participants. The worthy Chair IEEE Karachi Section, Prof. Dr. Bhawani Shankar Chowdhry, attended the workshop as the Chief Guest, delivered his speech, and appreciated the Speaker for taking their time from their very busy schedule. He also appreciated the efforts of Prof. Klara towards the research and activities related to IEEE ComSoc. The head of the Telecommunication department, MUET, Prof. Dr. Aftab Ahmed Memon, also attended this online talk as a guest of honor. The former head of the Telecommunication department MUET and currently working as a Professor, Prof. Dr. Faisal Karim Shaikh also attended this talk.



Talk delivered by Prof. Klara Nahrstedt, University of Illinois at Urbana Champaign, United States of America.



Welcome address by Dr. Umair Ahmed Korai.

The online talk last for about 90 minutes and there were quite good questions from the participants. Prof. Klara gave an excellent presentation about the Navigation Graphs for 360 Degrees Tile Based Video Streaming Systems. Team IEEE ComSoc Karachi Chapter received huge appreciation from Chair IEEE Karachi section, Prof. Dr. Bhawani Shankar and received excellent responses from participants. The talk was amazing and helpful for the students.

In last, the head of the Telecommunication department, MUET, Prof. Dr. Aftab Ahmed Memon gave the concluding remarks and vote of thanks to the chief guest, guest speaker, event organizers, collaborators, volunteers, and participants.

GCN GLOBAL COMMUNICATIONS NEWSLETTER

STEFANO BREGNI
Editor-in-Chief
Politecnico di Milano, Italy
Email: bregni@elet.polimi.it

FABRIZIO GRANELLI
Associate Editor
University of Trento, Italy
Email: fabrizio.granelli@unitn.it

IEEE COMMUNICATIONS SOCIETY – MEMBER AND GLOBAL ACTIVITIES
Ana Garcia-Armada, Vice-President for Member and Global Activities
Ricardo Veiga, Director for Member Services
Andres Navarro, Director of LA Region
Fawzi Behmann, Director of NA Region
Luca Foschini, Director of EMEA Region
Tomoaki Otsuki, Director of AP Region
Baek-Young Choi, Chair of the WICE Standing Committee
Periklis Chatzimisios, Chair of the YP Standing Committee

REGIONAL CORRESPONDENTS WHO CONTRIBUTED TO THIS ISSUE
Amit Patel, USA (a.j.patel@ieee.org)
Ewell Tan, Singapore (ewell.tan@ieee.org)

IEEE ComSoc IEEE Communications Society

www.comsoc.org/gcn
ISSN 2374-1082

Innovation in Constrained Codes

Kees A. Schouhamer Immink

The author has found application in all hard disk, non-volatile memories, and optical discs, such as CD, DVD, and Blu-Ray Disc, and they are now projected for usage in DNA-based storage.

Abstract

Constrained coding is a somewhat nebulous term which we may define by either inclusion or exclusion. A constrained system is defined by a constrained set of "good" or "allowable" sequences to be recorded or transmitted. Constrained coding focuses on the analysis of constrained systems and the design of efficient encoders and decoders that transform arbitrary user sequences into constrained sequences.

Constrained coding has extensively been used since the advent in the 1950s of digital storage and communication devices. They have found application in all hard disk, non-volatile memories, and optical discs, such as CD, DVD, and Blu-Ray Disc, and they are now projected for usage in DNA-based storage. We survey theory and practice of constrained coding, tracing the evolution of the subject from its origins in Shannon's classic 1948 paper to present-day applications in DNA-based data storage systems.

Introduction

Data are sent or stored over a communications channel and received in a distorted form, for example, by noise, which may lead to errors. Noise, albeit always present, is not the only source of malfunctioning.

As a typical example, we take a look at a very ancient form of communications, namely smoke signals (Fig. 1). The experienced scouts among the readers know that a sender creates puffs of smoke using a fire and a blanket. The smoke signals must be observed by the receiver and therefore preferably the sender stands atop a hill or mountain. The smoke channel is binary, that is, a puff or the absence of a puff, so let us say a puff of smoke denotes a 1 and the absence of a puff denotes a 0. Clearly, random noisy sources such as wind, rain, and fog may impair the quality of the reception of the puffs sent, and a receiver may have difficulty in deciphering the message sent.

There is another, fundamental, issue: a long run of 0s will burn the blanket, leading to catastrophic failure. I know this smoke signaling example probably sounds a little farfetched for electronics engineers, but in 1990 in the area of optical disc recording (also known as "burning"), a patent was granted that offers a solution to a problem, where "In order to extend the life of the laser, it is desirable to select a code such that the number of laser pulses needed for recording is minimized [1]."

Quoting from Shannon's paper [2]: A *constrained channel* or *input-restricted noiseless channel*

> is not capable to transmit all possible signals; certain sequences only may be allowed. These will be possible signals for the channel.

A *constrained code* implements the constraints by converting user data into "allowed" signals, where disallowed sequences of data are avoided. A constrained code further caters for receiver timing on bit, codeword, and frame level. Constrained coding has extensively been used since the advent in the 1950s of electronic storage and communication devices [3, 4].

We present *capita selecta* of constrained codes, operating or in the offing. We start with a historical note on Shannon's Figure 2, followed by a description of runlength limited codes. We detail energy harvesting constraints and balanced codes. We present an overview of constrained coding for DNA-based storage.

For reasons of space, we could not discuss all new applications of constrained sequences; for example, Pearson codes, which are used to facilitate detection in the presence of noise and unknown gain and offset [5]. Constrained codes can also be exploited to guide deep-learning-based detection of resistive random access memory [6]. Two-dimensional, chessboard-like, constraints are important for avoiding "sneaky" paths in non-volatile memories, such as *memristors* [7–10].

Shannon's "Figure 2"

In a typical data storage or communications device, we may distinguish two kinds of codes, namely codes for correcting errors and constrained codes for preventing errors (we ignore crypto and data compression codes as they are outside the scope). The well-known "Figure 1" in Shannon's paper [2] describes the configuration of coding techniques in a communications system (note that storage and retrieval of digital information is seen as a special case of digital communication [11]). Later, Shannon introduces the *input-restricted noiseless channel*, which we currently denote as a constrained channel. Shannon's less well-known "Figure 2" shows a graphical representation of the constraints on adjacent telegraph dots, dashes, and letter and word spaces. Figure 2 illustrates that a letter space may not follow a word space and *vice versa*. In a similar vein as Shannon's telegraphy example, we may

Digital Object Identifier:
10.1109/MCOM.002.2200249

The author is with Turing Machines Inc, The Netherlands.

formulate constraints for the above smoke channel. Say, after a given number of consecutive 0's (non-puffs), we must cool down the blanket by letting off some steam puffs.

Since specific signals are forbidden, we must define the data transmission capacity of such constrained channels. The *capacity* of a constrained channel is defined by [2]

$$C = \lim_{T\to\infty} \frac{1}{T} \log_2 N(T), \tag{1}$$

where $N(T)$ denotes the number of allowed signal sequences in the time interval t. The capacity sets an upper limit on the average amount of data that can be transmitted in a given time interval. The design of constrained codes entails implementing the channel constraints in an efficient way (i.e., as close to capacity as possible).

It has often been felt that rigorously forbidding the usage of certain vexatious sequences does not always lead to the best technical optimum. *Weak* constraints [12, 13] are more forgiving as they allow the overruling of some specified constraints with a prescribed (small) probability. As a result, the channel capacity will increase.

Runlength-Limited Codes

Runlength-limited (RLL) codes have been widely applied in magnetic and optical recording systems [14, 15], inclusive of CD, DVD, and BluRay Disc [16]. The number of consecutive 0s or 1s is called a *runlength*. For certain applications, we demand that the runlengths of both the 1s and 0s of the sent message lie between a minimum and maximum value. Note that the smoke signaling discussed in the Introduction has a maximum runlength constraint on the 0s (absence of a puff) only. We take a closer look at the counting of such sequences.

Let $\boldsymbol{x} = (x_1, ..., x_n)$ denote a sequence of n binary symbols $x_i \in \{0, 1\}$. A dk sequence satisfies two constraints, namely

d constraint: Consecutive 1s are separated by at least d 0's.
k constraint: Any run of consecutive 0s is limited to k.

Clearly, $k > d$. A dk sequence is translated into a binary RLL sequence by a simple precoding step. The RLL sequence, denoted by $\boldsymbol{y} = (y_1, ..., y_n)$ is obtained by $y_i = y_{i-1} \oplus x_i$, $y_0 = 1$, where $\oplus$ denotes mod 2 addition. The runlengths of the RLL sequence so obtained lie between $d + 1$ and $k + 1$. For example, let $\boldsymbol{x} = (0, 1, 0, 0, 1, 0, 0, 0, 1)$ be a dk sequence; then the associated RLL sequence obtained by mod 2 addition is $\boldsymbol{y} = (1, 0, 0, 0, 1, 1, 1, 1, 0)$.

In order to compute the capacity of the d-constrained channel, using Eq. 1, we need an expression for the number of d sequences of length n, denoted by $N_d(n)$, where for clerical reasons we drop the k constraint. For large n, the number of d sequences grows exponentially with length n, or $N_d(n) \approx c\lambda^n$, $n \gg 1$, where c is a constant and λ the growth factor [2]. The growth factor λ, $1 \le \lambda \le 2$, is the largest real root of the *characteristic equation*

FIGURE 1. Early example of a constrained communication system. A long run of 0s (absence of smoke puffs) will burn the blanket, destroying the transmitter (Image: uk.pinterest.com/pin).

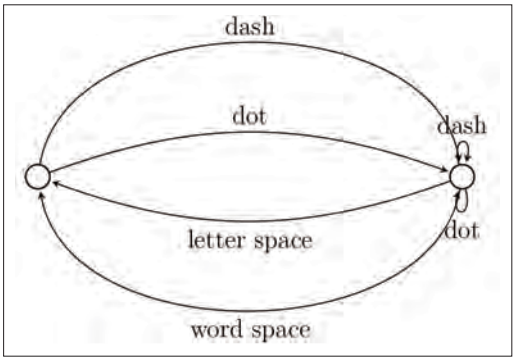


FIGURE 2. Shannon's "Figure 2" displays the channel constraints of a telegraphy signal. A letter space may not follow a word space and vice versa..

d	C_d
1	0.694
2	0.551
3	0.465
4	0.406

TABLE 1. Capacity C_d vs. minimum runlength d.

$$z^{d+1} - z^d - 1 = 0.$$

The channel capacity, denoted by C_d , is simply (refer to Eq. 1)

$$C_d = \log_2 \lambda.$$

Table 1 shows results of computations. We may note that the larger d, the smaller the capacity C_d.

There is an alternative technique to derive the channel capacity, presented by Shannon, which is based on the representation of the (dk) constraints by a finite-state sequential machine. A d-constrained channel, for example, can be modeled with $d + 1$ states (Fig. 3). A trajectory following the arrows and reading off the labels generates a d sequence. The finite-state machine

The capacity sets an upper limit to the average amount of data that can be transmitted in a given time interval. The design of constrained codes entails implementing the channel constraints in an efficient way (i.e., as close to capacity as possible).

Not all electronic products can easily be plugged into a power strip or can use batteries. Energy harvesting offers an alternative power source, where the receiving device reuses the energy carried by the received signals without the need for battery maintenance [21, 22].

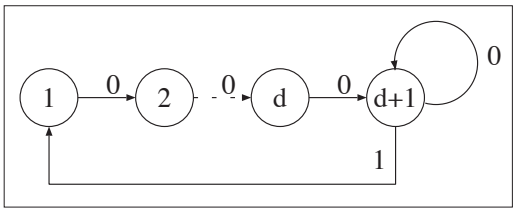


FIGURE 3. State-transition diagram for a (d) sequence. Only in state σ_{d+1} either a 0 or a 1 is emitted; in all other states, a 0 is emitted.

source	*output*
000	00000
001	00001
010	00010
011	00100
100	00101
101	01000
110	01001
111	01010

TABLE 2. Simple ($d = 1$) block code.

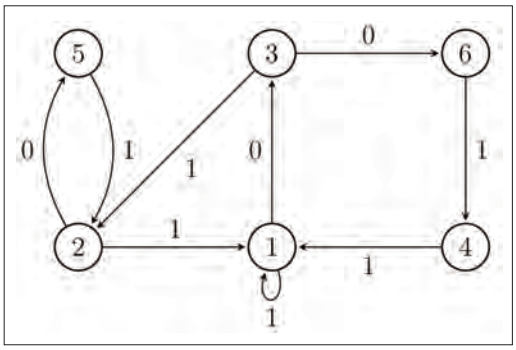


FIGURE 4. Stepping from state to state, following the direction of the arrows, and reading off the symbols tagged to the arrows generates a sequence with at least $t = 2$ 1s in a window of $\ell = 4$ bits.

model allows us to compute the capacity, and it may guide the code design [17, 18].

Implementation of RLL Codes

On the back of an envelope, we can easily write down all $d = 1$ sequences of four bits. There are exactly eight such words. The eight codewords, tabulated in the right-hand column of Table 2, are found by adding one leading 0 to the eight 4-bit words. Then, as a result, consecutive codewords can be freely cascaded, forming a very long sequence of symbols without violating the minimum $d = 1$ constraint. The first column shows 3-bit source words that are paired to 5-bit codewords. The quotient of the numbers of source bits and output bits, $R = 3/5$, is called the *rate* of the code, and denoted by R. The *code efficiency* is the quotient of code rate and capacity of the (d)-constrained channel having the same runlength constraints, that is, $0.6/0.694 \simeq 0.86$. Thus, the simple 5-bit look-up code considered above is sufficiently powerful to attain 86 percent of the rate that is maximally possible, which, actually, demonstrates that good efficiencies are feasible with very simple constructions. Note that since the first 0 symbol of the codeword does not carry information, it can be skipped by the receiver.

A maximum runlength, k constraint, can be implemented in the ($d = 1$) block code by noting that the first bit, which is preset to 0, can be set to a 1 if the last symbol of the preceding codeword and the second symbol of the current codeword to be conveyed are both 0; then the first codeword symbol can be set to 1 without violating the $d = 1$ channel constraint. This extra "merging" rule, which governs the selection of the first symbol, can be implemented with some extra hardware. It is readily conceded that with this additional rule, the ($d = 1$) code, presented in Table 2, turns into a ($d = 1$, $k = 6$) code.

It is straightforward to generalize the preceding ($d = 1$) block code to encoder constructions that generate sequences with a larger value of the minimum runlength. Choose some appropriate codeword length, and write down all d-constrained words that start with d 0s. Block codes based on a simple one-to-one relationship between source and codeword symbols can easily be generalized for other values of d and implemented. It has been found that *state-dependent* encoding may often lead to both simpler code implementation and a high rate efficiency [17, 19]. Large lookup tables can be avoided by enumerative methods [20].

Energy-Harvesting Constrained Codes

Not all electronic products can easily be plugged into a power strip or can use batteries. Energy harvesting offers an alternative power source, where the receiving device reuses the energy carried by the received signals without the need for battery maintenance [21, 22]. Such devices are anticipated in products of the Internet of Things (IoT). For binary systems emitting 0 and 1 signals, this has a bearing on the number of 1s (the 1s are supposed to carry the energy) that are sent in a prescribed time slot. A given number of 1s in transmitted signals is required to carry sufficient energy within a prescribed time span.

Two approaches have been studied for simultaneous information and energy communication, namely the *subblock-energy constraint* [23] and the *sliding-window constraint* [22]. A subblock-energy constraint guarantees a prescribed number of 1s per block. A sliding-window (ℓ, t)-constraint is satisfied if the number of 1s within any window of ℓ consecutive bits of that sequence is at least t. We can set up a finite-state machine that describes the (ℓ, t) constraint. For the case $\ell = 4$, $t = 2$, we obtain the 6-state finite-state machine shown in Fig. 4 [24]. Its capacity equals $C = 0.778$.

Spectrum Shaping Codes

Due to ac-coupling transformers, early digital communication systems were not able to transfer very low frequencies, and in order to avoid detrimental intersymbol interference, *dc-free* codes were used [3]. A *spectral null code* is a generic term for a code whose spectral content of long sequences at specified frequencies is nil or very small. The spectral null frequencies of primary interest in communications are the zero frequency (dc) and the Nyquist frequency. Dc-free codes have found widespread application in various fields such as data transmission and data storage [25–29].

Let $\boldsymbol{x}$ be an n-bit word of bipolar symbols, $x_i \in \{-1, 1\}$. The *running digital sum*, denoted by z_i, is defined by $z_i = z_{i-1} + x_i$, $z_0 = 0$. It has been shown by Justesen [30] that a sequence is dc-free if the running digital sum is bounded, meaning

that the long-term relative difference between the numbers of 1s and −1s transmitted is nil. Lookup tables for translating source words into sequences with limited running digital sum have been widely applied. For example, rate 8/10, 8b10b codes that translate eight source bits into ten channel bits to achieve bounded running sum, have been applied in communications [31] and magnetic tape recording (DCC and RDAT) [15]. The more efficient 64b66b dc-free code has been applied in modern fibre transmission [32].

Balanced Codes

A bipolar sequence, $\boldsymbol{x}$, of n bits, n even, is said to be balanced if the sequence has equal numbers of 1s and −1s. Knuth [33, 34] showed that by inverting a first segment of ℓ symbols, $1 \le \ell \le n$, any source word $\boldsymbol{x}$ can be balanced. The encoded balancing index ℓ is sent as a tag, requiring around $\log_2 n$ redundant bits so that the receiver can undo the modification. Knuth's implementation of balanced codes is attractive for encoding large source words as its complexity scales linearly with word length n, but it requires a redundancy $\log_2 n$, which is, for large n, around twice the minimum redundancy of a code comprising the full set of balanced codewords. Modifications of Knuth's generic scheme bridging the gap between the minimum redundancy and that of Knuth's implementation are discussed in [35].

Constrained Coding for DNA-Based Storage

Essentially, all digital archives are now stored on magnetic and optical media, but with the gigantic increase in data volumes, traditional storage media may rapidly run out of steam, and alternative solutions have to be sought. *Deoxyribonucleic acid*, DNA in short, is a polymer composed of two polynucleotide chains that coil around each other forming a double helix. DNA-based data storage, which stores information in synthetic strands of DNA, has three key advantages. It has extremely high data storage density, remains stable for hundreds of years, and requires very little power [36, 37]. Of course, there are a few engineering obstacles to be conquered before DNA can be used as a mass data storage device [38].

Naturally occurring DNA consists of four types of nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). A DNA strand (or oligo in short) is a linear sequence of these four nucleotides that are composed by DNA synthesizers. Binary source data are translated into the four types of nucleotides, for example, by mapping two binary sources into a single nucleotide. To retrieve the data, the synthetic DNA molecules are sequenced, and the output translated back into the original digital information.

It has been found that when the same nucleotide is repeated more than, say, four times in a row, the probability of sequencing errors rises substantially [39]. Also, oligos with large imbalance between GC and AT content are prone to read errors, and should therefore be avoided. Blawat's format [37] shows a constrained code that uses a lookup table for translating binary source data into strands of nucleotides with a runlength of at most three. In Erlich and Zielinski's experiments [40], both the maximum run requirement and the weak balance constraint are taken into account. Constrained codes that avoid both the maximum run requirement and the weak balance constraint can readily be designed with earlier theory developed in [29].

The requirements on constrained codes are much more involved than the "ordinary" balance and runlength constraints: a pool of DNA is like tea leaves in a tea pot, with minuscule free-floating molecules. This makes data restoration a coding challenge as the data does not have a prescribed place as in traditional storage devices. Constrained index codes make it possible to re-assemble the stored data from the free-floating strands [41–43].

Conclusions

We have presented an overview of innovation in constrained codes used in early telecommunications, mass data storage products, and future products such as energy-harvesting systems and DNA-based storage.

The requirements on constrained codes are much more involved than the "ordinary" balance and runlength constraints: a pool of DNA is like tea leaves in a tea pot, with minuscule free-floating molecules. This makes data restoration a coding challenge as the data does not have a prescribed place as in traditional storage devices.

References

[1] J. J. Verboom *et al.*, "Method of and Device for Recording Information, Record Carrier, and Device for Reading the Recorded Information," U.S. patent 4,930,115, May 1990.
[2] C. E. Shannon, "A Mathematical Theory of Communication," *Bell Sys. Tech. J.*, vol. 27, July 1948, pp. 379–423. DOI: 10.1145/584091.584093.
[3] K. W. Cattermole, *Principles of Pulse Code Modulation*, Iliffe Books Ltd, 1969.
[4] K. A. S. Immink, P. H. Siegel, and J. K. Wolf, "Codes for Digital Recorders," *IEEE Trans. Info. Theory*, vol. IT-44, no. 6, Oct. 1998, pp. 2260–99.
[5] R. Bu, *Coding Techniques for Noisy Channels with Gain and/or Offset Mismatch*, Ph.D. thesis, TU Delft, Dec. 6, 2021.
[6] X. Zhong *et al.*, "Constrained Coding and Deep Learning Aided Threshold Detection for Resistive Memories," *IEEE Commun. Letters*, vol. 26, no. 4, Apr. 2022, pp. 803–07. DOI: 10.1109/LCOMM.2022.3148292.
[7] R. Talyansky, T. Etzion, and R. M. Roth, "Efficient Code Construction for Certain Two-Dimensional Constraints," *IEEE Trans. Info. Theory*, vol. IT-45, no. 2, Mar. 1999, pp. 794–99. DOI: 10.1109/18.749031.
[8] E. Ordentlich and R. M. Roth, "Low Complexity Two-Dimensional Weight-Constrained Codes," *IEEE Trans. Info. Theory*, vol. 58, no. 6, June 2012, pp. 3892–99. DOI: 10.1109/TIT.2012.2190380.
[9] F. Sala, K. A. S. Immink, and L. Dolecek, "Error Control Schemes for Modern Flash Memories: Solutions for Flash Deficiencies," *IEEE Consumer Electronics Mag.*, vol. 4, no. 1, Jan. 2015, pp. 66–73. DOI: 10.1109/MCE.2014.2360965.
[10] G. Song *et al.*, "Performance Limit and Coding Schemes for Resistive Random-Access Memory Channels," *IEEE Trans. Commun.*, vol. 69, no. 4, Apr. 2021, pp. 2093–2106. DOI: 10.1109/TCOMM.2021.3051413.
[11] E. R. Berlekamp, "The Technology of Error-Correcting Codes," *Proc. IEEE*, vol. 68, no. 5, May 1980, pp. 564–93. DOI: 10.1109/PROC.1980.11696.
[12] M. Jin, K. A. S. Immink, and B. Farhang-Boroujeny, "Design Techniques for Weakly Constrained Codes," *Trans. Commun.*, vol. 51, no. 5, May 2003, pp. 709–14. DOI: 10.1109/TCOMM.2003.811377.
[13] D. J. C. MacKay, "Almost-Certainly Runlength-Limiting Codes," B. Honary, Ed., *Cryptography and Coding 2001*, LNCS, vol 2260. Springer. DOI: org/10.1007/3-540-45325-3-13.
[14] K. A. S. Immink, "Runlength-Limited Sequences," *Proc. IEEE*, vol. 78, no. 11, Nov. 1990, pp. 1745–59. DOI: 10.1109/5.63306.
[15] K. A. S. Immink, *Codes for Mass Data Storage Systems*, 2nd ed., Shannon Foundation Publishers, 2004; https://www.researchgate.net/publication/239666257. ISBN 90-74249-27-2.
[16] K. A. S. Immink, "A Survey of Codes for Optical Disk Recording," *IEEE JSAC*, vol. 19, no. 4, Apr. 2001, pp. 756–64. DOI: 10.1109/49.920183.
[17] R. L. Adler, D. Coppersmith, and M. Hassner, "Algorithms for Sliding Block Codes — An Application of Symbolic Dynamics to Information Theory," *IEEE Trans. Info. Theory*, vol. IT-29, no. 1, Jan. 1983, pp. 5–22. DOI: 10.1109/TIT.1983.1056597.
[18] P. A. Franaszek, "A General Method for Channel Coding,"

IBM J. Res. Develop., vol. 24, no. 5, 1980, pp. 638–41. DOI: 10.1147/rd.245.0638.
[19] B. H. Marcus, P. H. Siegel, and J. K. Wolf, "Finite-State Modulation Codes for Data Storage," *IEEE JSAC*, vol. 10, no. 1, Jan. 1992, pp. 5–37. DOI: 10.1109/49.124467.
[20] A. Hareedy, B. Dabak, and R. Calderbank, "The Secret Arithmetic of Patterns: A General Method for Designing Constrained Codes Based on Lexicographic Indexing," *IEEE Trans. Info. Theory*, 2022. DOI: 10.1109/TIT.2022.3170692.
[21] P. Popovski, A. M. Fouladgar, and O. Simeone, "Interactive Joint Transfer of Energy and Information," *IEEE Trans. Commun.*, vol. 61, no. 5, May 2013, pp. 2086–97. DOI: 10.1109/TCOMM.2013.031213.120723.
[22] E. Rosnes, A. I. Barbero, and O. Ytrehus, "Coding for Inductively Coupled Channels," *IEEE Trans. Info. Theory*, vol. 58, no. 8, Aug. 2012, pp. 5418–36. DOI: 10.1109/TIT.2012.2201370.
[23] A. Tandon, H. M. Kiah and M. Motani, "Bounds on the Size and Asymptotic Rate of Subblock-Constrained Codes," *IEEE Trans. Info. Theory*, vol. 64, no. 10, Oct. 2018, pp. 6604–19. DOI: 10.1109/TIT.2018.2864137.
[24] K. A. S. Immink and K. Cai, "Properties and Constructions of Energy-Harvesting Sliding-Window Constrained Codes," *IEEE Commun. Letters*, vol. 24, no. 9, Sept. 2020, pp. 1890–93, DOI: 10.1109/LCOMM.2020.2993467.
[25] D. T. Dao, H. M. Kiah, and T. T. Nguyen, "Average Redundancy of Variable-Length Balancing Schemes à la Knuth," ArXiv: 2204.13831, Apr. 2022.
[26] O. P. Babalola and V. Balyan, "Efficient Channel Coding for Dimmable Visible Light Communications System," *IEEE Access*, vol. 8, 2020, pp. 215,100–06. DOI: 10.1109/ACCESS.2020.3041431.
[27] J. N. Franklin and J. R. Pierce, "Spectra and Efficiency of Binary Codes without DC," *IEEE Trans. Commun.*, vol. COM-20, no. 6, Dec. 1972, pp. 1182–84. DOI: 10.1109/TCOM.1972.1091308.
[28] F. Chang *et al.*, "Design and Implementation of Anti Low-Frequency Noise in Visible Light Communications," *2017 Int'l. Conf. Applied System Innovation*, Sapporo, 2017, pp. 1536–38. DOI: 10.1109/ICASI.2017.7988219.
[29] K. A. S. Immink and K. Cai, "Properties and Constructions of Constrained Codes for DNA-Based Data Storage," *IEEE Access*, vol. 8, 2020, pp. 49,523–31. DOI: 10.1109/ACCESS.2020.2980036.
[30] J. Justesen, "Information Rates and Power Spectra of Digital Codes," *IEEE Trans. Info. Theory*, vol. IT-28, no. 3, May 1982, pp. 457–72. DOI: 10.1109/TIT.1982.1056516.
[31] A. X. Widmer and P. A. Franaszek, "A DC-Balanced, Partitioned-Block, 8B/10B Transmission Code," *IBM J. Res. Develop.*, vol. 27, no. 5, Sept. 1983, pp. 440–51. DOI: 10.1147/rd.275.0440.
[32] K. Balasubramanian, S. S. Agili, and A. Morales, "Encoding and Compensation Schemes Using Improved Pre-Equalization for the 64B/66B Encoder," *2012 IEEE Int'l. Conf. Consumer Electronics*, Las Vegas, NV, 2012, pp. 361–63. DOI: 10.1109/ICCE.2012.6161902.
[33] D. E. Knuth, "Efficient Balanced Codes," *IEEE Trans. Info. Theory*, vol. IT-32, no. 1, Jan. 1986, pp. 51–53. DOI: 10.1109/TIT.1986.1057136.
[34] F. Paluncic, B. T. Maharaj, and H. C. Ferreira, "Variable- and Fixed-Length Balanced Runlength-Limited Codes Based on a Knuth-Like Balancing Method," *IEEE Trans. Info. Theory*, vol. IT-65, no. 11, Nov. 2019, pp. 7045–66, DOI: 10.1109/TIT.2019.2914205.
[35] K. A. S. Immink and J. H. Weber, "Very Efficient Balanced Codes," *IEEE JSAC*, vol. 28, no. 2, Feb. 2010, pp. 188–92, DOI: 10.1109/JSAC.2010.100207.
[36] G. M. Church, Y. Gao, and S. Kosuri, "Next-Generation Digital Information Storage in DNA," *Science*, vol. 337, no. 6012, 2012, pp. 1628–28, DOI: 10.1126/science.1226355.
[37] M. Blawat *et al.*, "Forward Error Correction for DNA Data Storage," *Int'l. Conf. Computational Science 2016*, vol. 80, 2016, pp. 1011–22; doi.org/10.1016/j.procs.2016.05.398.
[38] D. Panda *et al.*, "DNA as A Digital Information Storage Device: Hope or Hype?," *Biotech*, no. 5, 2018, p. 239. DOI:10.1007/s13205-018-1246-7
[39] J. Bornholt *et al.*, "A DNA-Based Archival Storage System," *ACM SIGOPS Operating Systems Review*, vol. 50, 2016, pp. 637–49.
[40] Y. Erlich and D. Zielinski, "DNA Fountain Enables A Robust and Efficient Storage Architecture," *Science*, vol. 355, Mar. 2017, pp. 950–54.
[41] J. Sima, N. Raviv, and J. Bruck, "Robust Indexing - Optimal Codes for DNA Storage," *2020 IEEE Int'l. Symp. Info. Theory*, 2020, pp. 717–22. DOI: 10.1109/ISIT44484.2020.9174447.
[42] A. Lenz *et al.*, " Coding Over Sets for DNA Storage," *IEEE Trans. Info. Theory*, vol. 66, no. 4, 2020, pp. 2331–51; http://dx.DOI.org/10.1109/TIT.2019.2961265.
[43] I. Shomorony and R. Heckel, "Information-Theoretic Foundations of DNA Data Storage," *Foundations and Trends in Commun. and Info. Theory*, vol. 19, no. 1, 2022, pp 1–106; http://dx.DOI.org/10.1561/0100000117.

Biography

Kees A. Schouhamer Immink [M'81, SM'86, F'90] (immink@turing-machines.com) founded Turing Machines Inc in 1998. He was from 1994 to 2014 an adjunct professor at the Institute for Experimental Mathematics, Essen-Duisburg University, Germany. He contributed to digital video, audio, and data recording products, including Compact Disc, CD-ROM, DCC, DVD, and Blu-ray Disc. He received the 2017 IEEE Medal of Honor, a Knighthood in 2000, a personal Emmy award in 2004, the 1999 AES Gold Medal, the 2004 SMPTE Progress Medal, the 2014 Eduard Rhein Prize for Technology, and the 2015 IET Faraday Medal. He received the Golden Jubilee Award for Technological Innovation of the IEEE Information Theory Society in 1998. He was inducted into the Consumer Electronics Hall of Fame, and elected into the Royal Netherlands Academy of Sciences and the U.S. National Academy of Engineering. He received an honorary doctorate from the University of Johannesburg in 2014. He served the profession as President of the Audio Engineering Society inc., New York, in 2003.

CALL FOR PAPERS

IEEE COMMUNICATIONS MAGAZINE

FEATURE TOPIC: DATA SETS FOR MACHINE LEARNING IN WIRELESS COMMUNICATIONS AND NETWORKS

BACKGROUND

In the era of 5G and 6G, the telecommunication industry has evolved to provide services for new types of Internet of Things (IoT) devices in addition to basic mobile phone or internet services, including extended reality devices, sensors, and ground and aerial robots, to name a few. With the deployment of these new services, it is difficult for the wireless network to support ubiquitous connections with diverse quality-of-service (QoS) requirements. Despite the remarkable success of the model-based design and analysis for wireless networks, it turns out that they are not always adequate for dynamic wireless environments with diverse QoS requirements. To address this, the data-driven methods of Machine Learning (ML) is expected to enable fundamentally new intelligent design and decision-making in wireless networks. The intelligence supported by ML solutions will allow for the pattern recognition from time-series data, the network anomalies detection and prediction, the network design automation, and performance optimization in real-time, hence creating a self-optimizing and self-updating networks.

However, the research advancements with ML for wireless communications and networks inherently relies on the availability of data sets to test the results and attempts generalizations. One of the main bottlenecks for such research is the current limited availability of datasets from either practical simulations or experimental testbeds that can be considered as reference or standard data sets. Creating standard reference datasets for research purposes, ranging from low-level physical layer measurements to telecommunication network analyses, is expensive, especially for research organizations, while the commercial datasets from telecommunication operators are mostly inaccessible. Thus, in this Feature Topic (FT), we aim to attract manuscripts that propose data sets for experimental testbeds and ML methods in wireless communications and networking, which have the potentiality to become reference or standard data sets for research purposes. These manuscripts should encourage growth in the use of experimental methods and the use and analysis of data sets.

All submissions must be based on high-quality research that has already been published or accepted in a peer-reviewed venue, and this must be clearly indicated. Thus, this submission does not need to verify the underlying research. Instead, a submission should focus on the description of the experimental setup and the value of the considered public available data set.

This FT submission must provide a description of the necessary lab equipment and the steps for performing a physical experiment. Experiments do not need to be described at the level of a user manual, but should have substantially more details than any previous work using the same setup and the level of detail should enable the work to be reproducible by someone who is not an expert in the area.

An original data set from the experiment must be provided. This data set should be of demonstrable value for the IEEE communications community. There should also be a brief example(s) of how to use the data, and this should also be novel (and in particular distinct from previous works that used the method). Data sets should be placed on IEEE DataPort with open access (open access uploads are currently free for IEEE members) and be available as part of the review process. Complementarily, data and code can be uploaded to CodeOcean.

AI/ML Data sets for:

- 5G/6G testbeds and trials
- Distributed AI/ML over communication networks
- Distributed multi-agent reinforcement learning
- Edge learning in wireless networks
- Federated learning and communications
- Integrated sensing and communication
- Intelligent reflecting surfaces
- Learn to transmit and receive
- Link layer
- MAC layer
- Mobility and network management
- Molecular networks
- Optical networks
- Over-the-air computation
- Physical layer
- Privacy and security issues
- Resource management and network optimization
- Semantic communications
- Wireless communications and networks to support AI/ML services

SUBMISSION GUIDELINES

Manuscripts should conform to the standard format as indicated in the Information for Authors section of the *IEEE Communications Magazine's* Manuscript Submission Guidelines. Please, check these guidelines carefully before submitting since submissions not complying with them will be administratively rejected without review.

All manuscripts to be considered for publication must be submitted by the deadline through Manuscript Central. Select the "FT–2220/Data Sets for Machine Learning in Wireless Communications and Networks" topic from the drop-down menu of Topic/Series titles. Please observe the dates specified here below noting that there will be no extension of submission deadline.

IMPORTANT DATES

Manuscript Submission Deadline: 31 October 2022
Decision Notification: 28 February 2023
Final Manuscript Due Date: 31 March 2023
Publication Date: First Quarter/Second Quarter 2023

GUEST EDITORS

Carlo Fischione (Lead Guest Editor)
KTH Royal Institute of Technology, Sweden
carlofi@kth.se

Marwa Chafii
New York University Abu Dhabi, UAE
marwa.chafii@nyu.edu

Yansha Deng
King's College London, UK
yansha.deng@kcl.ac.uk

Melike Erol-Kantarci
University of Ottawa and Ericsson, Canada
melike.erolkantarci@uottawa.ca

MOBILE COMMUNICATIONS AND NETWORKS


Wanshi Chen


Ilker Demirkol


Miraj Mostafa


Stefano Ruffini

This issue of Mobile Communications and Networks presents six articles covering various important and interesting aspects in wireless communications, including network capacity analysis in the presence of public safety related operations, shared fronthaul handling focusing on compression control, exploration of future-looking multi-tenanted radio communication systems, an overview of Wi-Fi 8 evolutions addressing new applications, investigation of power consumption for 5G base stations using an artificial neural network (ANN) architecture, and study of off-network communications in railway communications.

Disasters and the resulting emergencies are increasingly presenting situations where telecommunications infrastructure is often compromised, leading to diminished network capacity. The first article, "CPAWS: Cognitive Public Alerts to Wireless Subscribers for Enhancing Public Safety Operations During Emergencies," presents a framework that is able to predict and prevent mobile network overload/outage during emergencies using a mix of behavioral studies and machine learning tools. The efficacy of wireless emergency alerts (WEAs) is analyzed in terms of saving network bandwidth by reducing users' non-essential traffic across designated cellphone applications. Complementary access and alert control strategies ensure that the network load remains below the diminished available capacity during emergencies.

The emergence of centralized-RAN (Radio Access Network) has made fronthaul a more critical component for the current and evolving RAN. With the breakdown of base station to different units, continued antenna densification in 5G, and possibility of sharing single fronthaul interface for multiple cells, both capacity and latency are becoming critical requirements for fronthaul. Our second article, titling "Fronthaul Compression Control for Shared Fronthaul Access Networks," analyzes compression control to mitigate the new fronthaul requirements. It proposes new integral solutions for dynamic fronthaul compression control in shared fronthaul architectures. The focus is compression control strategies for multiple-cell/multiple-user scenarios sharing a common fronthaul link, where various methods for modulation data compression and scheduling can be used. The article also presents simulation results in favor of the proposed optimized modulation compression strategies.

Radio communication systems of the future are expected to move away from fixed infrastructure providers and static contracts, and towards multi-tenanted systems featuring actively negotiated terms of service. The third article "Distributed Trust and Reputation Management for Future Wireless Systems," introduces a trust-based framework to implement a massively shared, multi-tenant wireless communication system. The presented solution is based on a distributed, three-layer, and trust-based hardware sharing scheme among operators that overcomes the limitations of a single operator owned monolithic network.

The new user applications, such as AR (Augmented Reality)/VR (Virtual Reality), define ever more stringent requirements for the underlying networks, not only for data rates, but also for latency, the number of devices to be supported, etc. The new Wi-Fi technologies will need to satisfy such requirements. The possible ways to achieve these requirements are presented in the fourth article titled "Future Directions for Wi-Fi 8 and Beyond." The article first overviews the evolutions up until IEEE 802.11be, and provide future challenges and directions for the new Wi-Fi solutions.

Power consumption is an important aspect for 5G base stations. The fifth article, "Machine Learning and Analytical Power Consumption Models for 5G Base Stations," presents a data-driven and multi-carrier power consumption model. An ANN architecture for modelling and estimating the power consumption is provided, followed by a demonstration of its good accuracy using data collected from realistic commercial deployments. Subsequently, an analytic model is proposed to make the power control model analytically tractable, which can help generalize its usefulness for understanding and analyzing power consumption in realistic networks.

An interesting communication domain studied in this issue is the railway communications, specifically the direct (off-network) communication between the railway nodes without the network relaying. The sixth article titled "Off-Network Communications for Future Railway Mobile Communication Systems: Challenges and Opportunities" first lists such use cases, and then analyzes the possible technologies that could support such communications. Such use cases have been also studied by the International Union of Railways within its Future Railway Mobile Communication System (FRMCS) outline. The article provides a quantitative comparative analysis of these technologies through simulations, along with a list of open challenges.

We appreciate the authors for their timely works that helped us picking the combination of the above six forward-looking articles to address the needs and interest of wider audience we have.

Thanks to the reviewers for their usual silent contributions in making sure that the high quality and strong relevance of the articles are not compromised. With multiple revisions, their efforts are amplified, but they still remain behind the mask. We also acknowledge the continuous support we have been enjoying from the editors and staff members.

We ask the readers to come forward and join us as an author or a reviewer to make it even wider effort in laying the steppingstone of future mobile communications and networking domain.

BIOGRAPHIES

WANSHI CHEN (wanshic@qti.qualcomm.com), [SM] is a senior director, Technology at Qualcomm Inc., where he is involved in 5G research and standardization. He is currently 3GPP TSG RAN plenary Chair appointed in April 2021. Previously, he was 3GPP TSG RAN WG1 Chair and successfully led the group to deliver both the first and second 5G releases on time and with high quality. The highest degree that he received is a Ph.D. degree in electrical engineering from the University of Southern California.

ILKER DEMIRKOL (ilker.demirkol@upc.edu) [SM] is an associate professor in the Department of Mining, Industrial and ICT Systems Engineering at the Universitat Politècnica de Catalunya, Barcelona, Spain, where his research currently is focused on network algorithmics, the Internet of Things, and mobile networks. Over the years, he has worked in a number of research laboratories and industrial corporations in Europe and the United States. He has co-authored more than 75 ACM/IEEE journal and conference papers, including the recipients of the Best Paper Award at IEEE ICC '13 and the Best Demo Award at IEEE MASS 2019.

MIRAJ MOSTAFA (miraj.mostafa@americantower.com), is a senior manager for Network Engineering at American Tower Corporation. Previously, he worked for Nokia and Microsoft. He also contributed to standardization and industry collaboration activities in 3GPP, IEEE 802.11, Wi-Fi Alliance, and GSM Association. He received his Ph.D. in communications engineering, Master of Engineering in telecommunications, and Bachelor of Science in EEE from Tampere University of Technology, Asian Institute of Technology, and Bangladesh University of Engineering and Technology respectively.

STEFANO RUFFINI (stefano.ruffini@calnexsol.com) graduated in telecommunication engineering from the University of Rome La Sapienza. After a long experience at Ericsson, he is currently working as Strategic Technology Manager at Calnex Solution. He is currently contributing to ITU-T SG15 Q13 (serving as Rapporteur), IEEE1588, and other relevant synchronization standardization bodies and forums. He has published several international journal papers and delivered talks at various conferences. He is the Chairman of the International Timing and Sync Forum and a member of the Workshop on Synchronization in Telecommunication Systems Executive Committee.

CALL FOR PAPERS

IEEE COMMUNICATIONS MAGAZINE

FEATURE TOPIC: SEMANTIC COMMUNICATIONS: TRANSMISSION BEYOND SHANNON

BACKGROUND

In contrast to the Shannon paradigm whose underlying principle is to guarantee the correct reception of each single transmitted packet regardless of its meaning, semantic communication is concerned with the problem of how transmitted symbols convey a desired meaning to the destination, as well as how effectively the received meaning affects the action in a desired way. By communicating the meaning or semantics of the data, semantic communication holds the promise of making wireless networks significantly more energy-efficient, robust and sustainable.

Advancements in artificial intelligence (AI) provide a powerful tool for solving fundamental problems in semantic communications. As a result, significant efforts have been made recently to design machine learning (ML)-based semantic communications for future wireless networks. To build a pathway to semantic communications, the system should be redesigned carefully.

The Feature Topic (FT) aims to focus on the theoretical analysis and implementation design for semantic communications. Prospective authors are invited to submit original manuscripts, either technical or tutorial articles, on topics including but not limited to:

- Semantic entropy
- Semantic sampling and quantization
- Semantic compression
- Semantic information pursuit for multimodal data
- Semantic coding and signal processing
- End-to-end semantic communication system design
- Multi-agent reinforcement learning for semantic communications
- Distributed learning architectures for semantic communications
- Semantic interference control
- Resource management for semantic communications
- Network architectures and protocols for semantic communications
- Privacy and security issues in semantic communications
- Efficient/scalable neural network architectures and training algorithms for semantic communications
- Experiments and testbeds for semantic communications
- Semantic communications in emerging wireless networks, i.e., virtual reality and autonomous driving

The authors of survey or tutorial articles are strongly encouraged to have research experience in the area of semantic communication evidenced by technical publications or inventions.

SUBMISSION GUIDELINES

Manuscripts should conform to the standard format as indicated in the Information for Authors section of the *IEEE Communications Magazine's* Manuscript Submission Guidelines. Please, check these guidelines carefully before submitting since submissions not complying with them will be administratively rejected without review.

All manuscripts to be considered for publication must be submitted by the deadline through Manuscript Central. Select the "FT-2218/ Semantic Communications: Transmission Beyond Shannon"" topic from the drop-down menu of Topic/Series titles. Please observe the dates specified here below noting that there will be no extension of submission deadline.

IMPORTANT DATES

Manuscript Submission Deadline: 15 November 2022
Decision Notification: 15 March 2023
Final Manuscript Due: 25 March 2023
Publication Date: Second Quarter 2023

GUEST EDITORS

Geoffrey Ye Li
Imperial College London, UK
Geoffrey.Li@imperial.ac.uk

Yonina Eldar
Weizmann Institute of Science, Israel
yonina.eldar@weizmann.ac.il

Xiaoming Tao
Tsinghua University, China
taoxm@tsinghua.edu.cn

Zhijin Qin
Queen Mary Univ. London, UK
z.qin@qmul.ac.uk

Arash Behboodi
Qualcomm Inc.
abehboodi@qualcomm.com

Guangyi Liu
China Mobile, China
liuguangyi@chinamobile.com

Yiqun Ge
Huawei, Canada
Yiqun.Ge@huawei.com

CPAWS: Cognitive Public Alerts to Wireless Subscribers for Enhancing Public Safety Operations During Emergencies

Mohammad Yousefvand, Demetrios Lambropoulos, and Narayan Mandayam

The authors study the efficacy of wireless emergency alerts (WEAs) in terms of saving network bandwidth by using surveys on Amazon MTurk to test the effectiveness of seven designed WEAs on reducing users' non-essential traffic across 12 designated cell phone applications.

Abstract

Disasters and the resulting emergencies are increasingly presenting situations where telecommunications infrastructure (e.g., functioning base stations) is often compromised, leading to diminished network capacity. In these situations, providing trapped populations in isolated areas with network access to be able to communicate with rescue and recovery personnel or volunteer teams is critical. Further, when network capacity is diminished, it is essential that mobile network users be alerted to "comply" and not overload the network with non-essential bandwidth-intensive traffic unrelated to the emergency unfolding. While the current evolution of the Integrated Public Alert and Warning System (IPAWS) is able to generate and disseminate such emergency alerts, there has been no analysis or evaluation of the effectiveness of such alerts as well as the design of mechanisms that can ensure such compliance to prevent network outage. In this work, we present a framework for Cognitive Public Alerts to Wireless Subscribers (CPAWS) that is able to predict and prevent mobile network overload/outage during emergencies using a mix of behavioral studies and machine learning tools that consider real-time monitoring information of the environmental status, network situation, and user compliance. Specifically, we study the efficacy of wireless emergency alerts (WEAs) in terms of saving network bandwidth by using surveys on Amazon MTurk to test the effectiveness of seven designed WEAs on reducing users' non-essential traffic across 12 designated cell phone applications. Additionally, we design complementary access and alert control strategies to ensure that the network load remains below the diminished available capacity during emergencies. CPAWS is also backward-compatible since it extends IPAWS by adding a feedback loop for real-time monitoring and control.

Introduction and Motivation

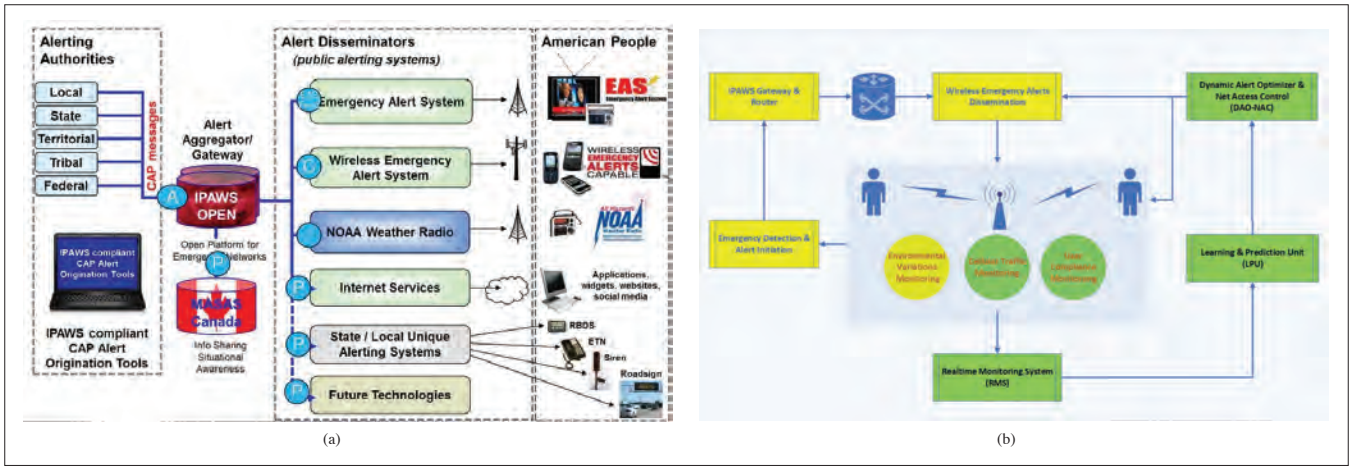
During emergencies and natural disasters, the mobile network capacity could dramatically shrink due to possible damages to communications infrastructure [1]. In these situations, providing trapped populations in isolated areas with network access to be able to communicate with rescue and recovery personnel or volunteer teams is critical [2]. Note that emergency personnel and first responders have their own dedicated channels and networks (e.g., FirstNet developed by AT&T) for communications. However, by preventing network outage and providing cellular access to mobile users (who cannot access dedicated channels), we can facilitate communications between users in need and rescue teams. In the current Integrated Public Alert and Warning System (IPAWS) [3], wireless emergency alerts (WEAs) can be sent to users in the affected areas to update them about the latest environmental conditions and to ask them to refrain from congesting the mobile network with non-essential traffic, which can result in network outage [4]. However, recent studies show that most mobile users will not fully comply with WEAs in terms of reducing non-essential cellular traffic [5]. Further, in the existing IPAWS, there is no monitoring mechanism to measure the effectiveness of WEAs and the associated user compliance levels. Also, IPAWS is not configured to dynamically react to real-time monitoring information and improve compliance [6].

To increase user compliance and provide network access for users who are in need of help, we present a framework for Cognitive Public Alerts to Wireless Subscribers (CPAWS), which is able to predict and prevent mobile network outage during emergencies using prediction models and machine learning (ML) tools that consider real-time monitoring information of the network, environment, and mobile user traffic [6]. With the advent of new app types with different priority levels in 5G networks including ultra-reliable low-latency communications (URLLC), it is possible to protect scarce network resources from non-essential use by decoupling low-priority non-essential user traffic from high-priority traffic using deep packet inspection techniques, or by checking the quality of service (QoS) flow ID (QFID) of each packet, allocated by the service data adaptation protocol (SDAP), and its associated priority, or by identifying the type of regular or priority uplink grant used for the transmission of each data packet [7]. The designed CPAWS architecture is able to protect scarce network resources from non-essential use by decoupling the low-priority non-essential user traffic from high-priority traffic that is critical for rescue operations. CPAWS is also

Digital Object Identifier:
10.1109/MCOM.001.2100897

Mohammad Yousefvand is with Qualcomm, USA; Demetrios Lambropoulos and Narayan Mandayamare with Rutgers University, USA.

0163-6804/22/$25.00 © 2022 IEEE



(a) (b)

FIGURE 1. a) Integrated Public Alert and Warning System (IPAWS) [3]; b) Cognitive Public Alerts to Wireless Subscribers (CPAWS).

backward-compatible with the current IPAWS architecture as it complements IPAWS by adding a cognitive control cycle to it by which the real-time monitoring information of the system will be utilized to make it more adaptive and efficient. The main objective of the CPAWS architecture is to provide network access and prevent network outage during emergencies, through:

1. Boosting users' compliance by disseminating more customized and effective WEAs
2. Using complementary admission control and resource allocation mechanisms to ensure the traffic load will not exceed the diminished available capacity, which can result in network outage

The findings of a human subject survey on WEA compliance and accompanying simulation results of network performance validate the necessity and efficiency of the proposed CPAWS architecture.

WEA Evolution and IPAWS Architecture

The initial version of WEA messages called WEA 1.0 supported the transmission of alerts across three different classes, namely presidential alerts, imminent threat alerts, and child abduction (AMBER) alerts. These alerts, which became functionally available after 2012 [4], trigger a unique sound and vibration on receiving smartphones. Initial WEA alerts were used for broadcasting tornado, flash flood, dust storm, hurricane, typhoon, extreme wind, and tsunami warnings by means of broadcasting 90-character text-only messages to large geographical areas that were believed to be at risk.

Motivated by studies in [8], improvements were made in WEA 2.0 [9] by increasing the text length to 360 characters; also, new features and for public safety were established such as alert message prioritization, transmission of embedded references, multimedia content, and Spanish language alerts. Finally, the latest version of such alerts, WEA 3.0, was proposed in May 2019 to better integrate WEA modules with other components in the IPAWS architecture, and define broadcast domains of WEA alerts more precisely by narrowing geo-targeting requirements [10]. For example, WEA 3.0 specifications are assumed to be met only when an alert is delivered to 100 percent of users in a target area with no more than 0.1 mi overshoot, to prevent coarse broadcasts and increase the effectiveness of alerts.

WEAs are disseminated to users via IPAWS, which is managed by the Federal Emergency Management Agency (FEMA) [3]. The architecture and the interdependencies between different modules within IPAWS are illustrated in Fig. 1a. As shown, the IPAWS architecture includes four vertical layers, namely alerting authorities, alert aggregator/gateway, alert disseminators, and people/end users. The first layer from the left specifies the alert initiator authority at the local, state, or national level. All the alerts initiated in the first layer will be routed to the centralized alert aggregator/gateway in the second layer, which is managed by IPAWS Open Platform for Emergency Networks (IPAWS OPEN). The responsibility of IPAWS OPEN is to classify and process received alerts based on their type and scope and forward them to the appropriate alert dissemination system in the third layer. The WEA system is one of the modules in the third layer of IPAWS architecture that broadcasts the received alerts from the IPAWS gateway to mobile devices located within the target area defined for each WEA alert via wireless communications media and links. The third layer of IPAWS is, in fact, the broadcast layer, which delivers different emergency alerts to users via different shared communications media including wired, wireless, radio, or satellite communications links depending on the alert type, scope, and required technology for their propagation and delivery. The fourth layer includes the target audience of different emergency alerts who are accessible via different media.

CPAWS

The IPAWS architecture is only able to deliver WEAs to users during emergencies; however, after delivery of such alerts, there is no monitoring mechanism in IPAWS to see if users are actually complying with received alerts, or if the alerts themselves are effective enough to enhance users' compliance and motivate them to follow the suggested guidelines [3]. Considering the detrimental effects of mobile user non-compliance on the rescue and recovery operations during emergencies, we present the novel CPAWS architecture, which can enhance the efficiency of WEAs and increase user compliance [6].

The functional architecture of the proposed CPAWS model and the interdependencies

CPAWS is an adaptive and self-reconfigurable system that takes into account the current status of the mobile network, environment, and users to optimize its performance and ensure network access for trapped people in affected areas and enables them to get connected to the first responders and volunteers in the area if needed.

between its modules are illustrated in Fig. 1b. CPAWS is an adaptive and self-reconfigurable system that takes into account the current status of the mobile network, environment, and users to optimize its performance and ensure network access for trapped people in affected areas and enables them to get connected to the first responders and volunteers in the area if needed. CPAWS is also backward-compatible and could easily be integrated with the current IPAWS architecture, since it extends IPAWS by adding a cognitive cycle to make it more effective and adaptive based on the real-time information of the mobile network, users, and the environment [6]. As shown in Fig. 1b, the CPAWS architecture includes several new modules to extend the IPAWS functionalities. The yellow modules capture the original functionalities of the IPAWS architecture, and the additional green modules, namely real-time monitoring system (RMS), learning & prediction unit (LPU), and dynamic alert optimizer & network access control (DAO-NAC), are the modules that introduce a cognitive cycle with the ability to dynamically optimize the alerts and perform network access control mechanisms, according to the real-time monitoring information and prediction data.

The RMS module dynamically monitors the system's status and captures the real-time variations in the environment, users' traffic, and mobile network load and capacity. It also classifies the observed data traffic into some designated application categories described later. In addition, it measures the users' compliance level and their reaction to received alerts to see how effective each alert is, and if users are following the suggested guidelines in the alerts. It also monitors the variations in the status of the underlying natural disaster like flooding or a hurricane, and sends all the raw monitoring data to the learning and prediction unit (LPU) for further processing. The LPU receives the raw monitoring data in three domains relating to:

1. The network
2. The environment
3. The users

It then feeds the data into models for predicting the future state of the system using machine learning tools. More specifically, it predicts if a mobile network outage is likely to happen in the near future based on the current information. Note that the precision of the predictions made by the LPU for predicting the future status depends on the availability of enough training data for the ML models used by the LPU, and we assume the monitoring information gathered by RMS can be fed into ML models for this purpose. The LPU can also predict the effect of each possible action, like changing the alert type or blocking access for a given application type, on the system's performance. After processing the raw monitoring data, it extracts useful information about the current and future status of the system and the effects of each possible action on systems performance. Finally, LPU sends all the extracted information to the DAONAC module for tailoring the alerts and optimizing network access [6].

DAO-NAC includes two complementary sub-modules, DAO and NAC. DAO is responsible for generating customized, informative, and effective alerts based on the information gathered on the performance of different alerts, network capacity, mobile load, and also demographic information of the population in the affected areas. Hence, for each group of users, a customized alert message will be chosen from a variety of persuasive alerts in order to maximize their compliance and reduce non-essential traffic. Note that having prior demographic information on the target audience of WEAs may help to boost their compliance; for example, if the primary spoken language of a given population is Spanish, we might customize the alerts to be sent in Spanish. Since compliance to alerts by itself may not be sufficient to meet traffic demands under diminished network capacity during disasters, the NAC sub-module implements network access control strategies to ensure that communication resources are available for use by those in dire need. Specifically, the NAC sub-module classifies the application types based on their priority, and if needed, it starts blocking access for users who are congesting the network with non-essential low-priority traffic until the load goes below a given threshold [6].

In summary, the cognition of this system comes from two capabilities:

1. Ability to predict the future state of the system and the possibility of network outage, based on the current monitoring information
2. Ability to pick the most effective alert for the target audience in each geographical area, and also enforce complementary network access control mechanisms in case the prediction models suggest that network outage is imminent

Furthermore, note that CPAWS is an extension of IPAWS, which is a U.S.-based alert system; however, the proposed cognitive cycle in CPAWS, formed by green modules in Fig. 1, could be generalized and utilized in any alert system that does not have such closed-loop feedback control and optimization modules.

The three major building blocks of CPAWS are the *monitoring* system, the *learning and prediction* system, and the *optimization* system, along with tools and enabling technologies for each of these. Monitoring includes network traffic analysis, environmental sensing (flood levels, damage to cell towers, etc.) via Internet of Things (IoT) sensors and devices, and active tracking of end-user behavior and compliance via comparative network load analysis before and after alerts. Learning and prediction includes data collection and use of ML algorithms (e.g., linear regression, logistic regression, support vector machines, classification, reinforcement learning) for traffic prediction, outage prediction, alert compliance prediction, and even predicting environmental conditions in disaster areas. Optimization includes using monitoring, learning, and prediction information to design the best choice of alert for a given geographical area and target audience in order to improve alert compliance and also save the network from outage via complementary access control mechanisms. Implementing CPAWS requires knowledge of the types of applications that mobile users from different demographic groups use in everyday life, and also their compliance behavior to different types of alerts [6]. To this end, we next present the results of a human subject study.

Survey on Designing WEAS

To increase mobile users' compliance with emergency guidelines, we need to design effective alerts. Specifically, we design seven different alerts with different objectives ranging from purely persuasive and *Altruistic* alerts to *Punitive* alerts, and find the most effective ones for each category of people/residents in each emergency situation. Specifically, we conducted a survey in Amazon Mechanical Turk (MTurk) and recorded the responses of 898 participants who were smartphone users, U.S. residents, and fluent English speakers. Among them, 51 percent were men and 49 percent were women, with ages ranging from 18 to 65+ with a majority in the ranges 25–34 (33.96 percent) and 35–44 (33.49 percent). Further details on users' demographic information, the recruitment process, and the survey procedures can be found in [12].

Designing Seven Different Emergency Alerts

The seven different emergency alerts designed are referred to as Basic Information, Altruistic, Multimedia, Negative Reinforcement, Positive Reinforcement, Reward, and Punitive alerts [12]. The wording of these seven alerts is given below:

- ***Basic Information***: This is an emergency alert asking you to facilitate rescue and recovery operations by not using your cell phone until further notice.
- ***Reward***: This is an emergency alert asking you to facilitate rescue and recovery operations by not using your cell phone until further notice. By not using your cell phone until further notice, you will get up to a 50 percent discount on your next bill depending on your refrain period.
- ***Punitive***: This is an emergency alert asking you to facilitate rescue and recovery operations by not using your cell phone until further notice. Our data shows that you are using bandwidth-intensive apps. If you do not refrain and continue using such apps, you may lose your network access for a few hours/days.
- ***Multimedia***: This is an emergency alert asking you to facilitate rescue and recovery operations by not using your cell phone until further notice. (This message is sent along with a picture of the underlying rescue operations.)
- ***Negative Reinforcement***: This is an emergency alert asking you to facilitate rescue and recovery operations by not using your cell phone until further notice. If you use your cell phone, you may contribute to the suffering and loss of lives for the victims.
- ***Positive Reinforcement:*** This is an emergency alert asking you to facilitate rescue and recovery operations by not using your cell phone until further notice. If you do not use your cell phone, you are personally assisting the efficient retrieval and aid of the victims.
- ***Altruistic:*** This is an emergency alert asking you to facilitate rescue and recovery operations by not using your cell phone until further notice. If you refrain from all cell phone use, you are personally assisting the efficient retrieval and aid of the victims.

Assuming a hypothetical emergency situation, we measure the compliance of users after receiving the above mentioned alerts. The *Basic Information* alert just passes the information about the emergency and asks users to not use their cell phones until further notice. The *Reward* and *Punitive* alerts would add a financial benefit or punishment, respectively, for complying or disregarding the alert to not use cell phones until further notice. The *Positive Reinforcement* alert reiterates the positive impact of complying with the alert to facilitate the rescue and recovery of victims. The Negative Reinforcement alert highlights the negative impacts of disregarding the alert, which could add to the suffering and result in loss of life. The Multimedia alert includes a picture of the rescue operations to garner more attention from users and increase their compliance. Finally, the Altruistic alert builds on the Positive Reinforcement alert by appealing to the inherent humaneness of the users by asking them to refrain from all cell phone use.

Apps Category\ Info	Data Rate (Kbps)	Bandwidth Consumption Rank	Net Access Priority Rank
Video Streaming	530	High	Low
Send/Upload Videos	180	High	Low
Social Media	80	High	Low
Video Call	70	High	Low
Audio Call	19.05	Medium	High
Music	14.51	Medium	Low
Send/Upload Photos	14.22	Medium	Low
Web Browsing	4.27	Medium	Medium
Gaming	3.41	Medium	Low
Other Phone Apps	2.67	Medium	Medium
GPS	0.79	Low	High
SMS	0.02	Low	High

FIGURE 2. 12 categories of applications used by mobile users.

Classifying Cell Phone Applications into 12 Categories

To estimate mobile users' daily data traffic, we classify cell phone applications in terms of their bandwidth consumption into 12 categories, and then query survey participants on their daily usage time of each application category. As shown in Fig. 2, we divide the 12 categories of mobile applications into three classes of high, medium, and low data rate applications corresponding to their expected bandwidth requirement as identified in mobile service provider reports (e.g., [11]). In order to design network access control mechanisms during emergencies, we characterize applications into three (high, medium, low) priority classes, as shown in Fig. 2.

Survey Results

Our survey had three primary objectives:

1. To test users' understanding of mobile applications' bandwidth consumption
2. To record and estimate average mobile users daily traffic consumption
3. To monitor and measure the effectiveness of each of the seven alert types designed to reduce cellular traffic during emergencies [12]

Our survey reveals that for objective (1), the majority of participants were unable to correctly classify applications based on their bandwidth requirements. This could very well be one of the reasons for the low compliance to the emergency alerts, as we see in the following. For objective (2), the mobile users' average daily usage time

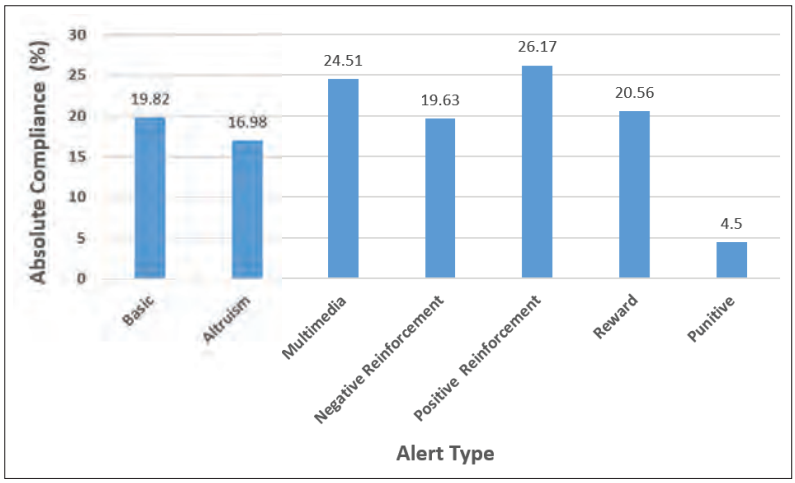


FIGURE 3. Absolute compliance across all apps.

per application is shown in [12], in which web browsing, listening to music, using social media, and video streaming are the most frequently used. Also, the mobile users' average daily traffic per application category could be found in our survey results [12], which is given by the product of the corresponding usage duration per application and the estimated data rate of that application.

For objective (3), we test the effectiveness of the designed alerts in reducing non-essential cellular traffic by randomly showing one of the designed alerts given earlier to each user, and asking about their cell phone applications usage during a hypothetical emergency situation after receiving the alert. To measure users' compliance with alerts, we define absolute compliance as a situation where users agree to stop using their cell phones until further notice. Partial compliance refers to a situation where users refrain from using a particular application category, while they could still use other applications on their cell phones. Figure 3 shows the absolute compliance of users with the received alerts across different application categories.

As shown in Fig. 3, the observed compliance with each of the seven designed alerts is different due to their different wording and motivation. However, the overall absolute compliance (across all application categories) is low, as on average only 18.88 percent of users would comply with the received alerts. There are several reasons that could explain the low compliance of users, as also observed in [13, 14]. In addition to the aforementioned lack of understanding of application bandwidth consumption, the other reasons for non-compliance could be:

1. The addiction of mobile users to cell phone usage in this era, which makes it extremely hard for them to abandon cell phone usage
2. Seeking news using cell phone applications, specifically social media applications, which are used by many users as a major source of getting news and updates during natural disasters
3. Lower *Altruistic* behavior in general. We also observe that more persuasive alerts (e.g., *Positive Reinforcement*) result in a higher compliance rate as compared to alerts that penalize (e.g., *Punitive*)

Although absolute compliance rates are low for almost all the alert types in Fig. 3, the partial compliance rates shown in [12] across different application categories are moderately higher, suggesting that even though users may stop using a specific application category, they still continue to use other applications. Consistent with earlier studies, the lowest partial compliance is observed for social media, audio calls, and web browsing, which are all applications that could be in high use during emergencies.

Regardless of the reasons for users' low compliance rates, it is clear that only relying on users' compliance would not be necessarily enough to save the network from outage as a result of excessive traffic. Hence, we need to implement traffic admission control policies at the network side as a complementary measure to maintain network access for mobile users during disasters [6]. In this work, we assume mobile service providers are able to dynamically monitor the traffic on their network and implement access control mechanisms if needed. In the next section, using simulation results, we show how we can use network access control mechanisms as a complementary measure with alerts to maintain the load below a given threshold, which is necessary to prevent network outage. Since IPAWS does not have any feedback control loop for alert optimization, it is associated with the basic alert, whose compliance performance is shown in Fig. 3.

Simulation Results

In this section, we present simulation results to show the impact of several parameters on the outage probability of cellular networks during emergencies. We evaluate the effectiveness of alerts in reducing non-essential traffic as well as the impact of network access control mechanisms to maintain the real-time traffic below a given threshold. We consider a cellular network model in which a cellular LTE base station located at the center of the cell is serving a varying number of mobile users who are randomly distributed within the cell area with a radius of 2 km. We assume the cellular LTE BS is reusing 20 MHz of licensed spectrum across its three sectors, and overall it is designed to provide the max capacity of 900 Mb/s, with 300 Mb/s capacity across each sector.

To estimate the cellular traffic at each time in our simulated network, we assume each active cellular user is randomly using one of the 12 categories of applications listed in Fig. 2, and also randomly receiving one of the seven alert types discussed above during a natural disaster, asking them to refrain from cell phone usage until further notice to facilitate rescue and recovery operations. We also assume the users' full and partial compliance levels are in accordance with the observed survey results presented in Fig. 3 and in [12]. We assume due to damages to cell tower infrastructure during the natural disaster, the cell capacity is diminished from 900 Mb/s to 300 Mb/s. Hence, to prevent an outage, we have to maintain the aggregate mobile traffic below this threshold.

The field measurements and cellular traffic patterns observed in residential areas in China presented in [15] show that the peak cellular traffic during daily busy hours is almost uniformly distributed due to a large number of users and central limit theorem, which suggests that the traffic will

converge to its average. Hence, in our simulated scenario, we assume the number of active mobile users is a Gaussian distributed random variable with a mean of 7000 users and standard deviation of 54.77 users, which makes the variance around 3000 users. Note that we assume the area's population is 10,000 users; however, on average, 7000 of them are active simultaneously during the peak traffic hours. Under such assumptions, Fig. 4 shows the variations of traffic over time and compares its fluctuation in four different scenarios:

1. No alert
2. Positive Reinforcement alert only
3. Punitive alert only
4. Complementary access control and Positive Reinforcement alert, enabled by CPAWS [6]

As seen, the aggregate traffic of users on normal days without alerts fluctuates between 500 Mb/s and 600 Mb/s, which is much higher than the available reduced capacity during a natural disaster. Also, we can see that although alerts could lower the total traffic depending on their effectiveness, ranging between Positive Reinforcement alert as most effective and Punitive alert as least effective, they are still not enough to lower the traffic below the required 300 Mb/s threshold to prevent network outage. The complementary access control mechanism along with the most effective alert, which is enabled by CPAWS, is able to maintain the total load below the required 300 Mb/s.

In our simulations, when access blocking needs to be enforced, we block the data traffic belonging to the low-priority applications, designated as gaming, music, social media, video streaming, and video uploads (according to Fig. 2) up until the congestion is resolved and the incoming traffic falls below the reduced capacity. Defining the traffic blocking rate as the ratio of excessive traffic (beyond the available capacity) to the total traffic, we can see its variations in Fig. 5 under three different scenarios:

1. No alert
2. Positive Reinforcement alert only
3. Complementary access control and Positive Reinforcement alert [6]

Since IPAWS does not have any access control mechanism, its performance is bounded by the performance of scenario 2. The number of active users is 10,000, and the diminished cellular capacities are 300 Mb/s, 500 Mb/s, and 800 Mb/s. As seen, when the capacity is diminished beyond a point, alert mechanisms by themselves are insufficient, and the CPAWS-enabled complementary access and alert control mechanism is necessary to prevent traffic blocking and provide access.

Conclusion

We present a framework for Cognitive Public Alerts to Wireless Subscribers (CPAWS) that is able to predict and prevent mobile network outage during emergencies using a mix of behavioral studies and machine learning tools that consider real-time monitoring information of the environmental status, network situation, and user compliance. Specifically, we study the efficacy of WEAs in terms of saving network bandwidth by using surveys on Amazon MTurk with seven designed WEAs on reducing users' non-essential traffic across 12 designated cell phone applications. In general, alerts that reinforced positive behavior were more effective than ones designed to be punitive. Despite the positive impact of alerts in reducing traffic, there is still a need to implement complementary access control to ensure that the cellular load remains below the diminished capacity; otherwise, network outage might be inevitable. We designed a complementary access and alert control mechanism to ensure that the network load always remains below the diminished available capacity during emergencies.

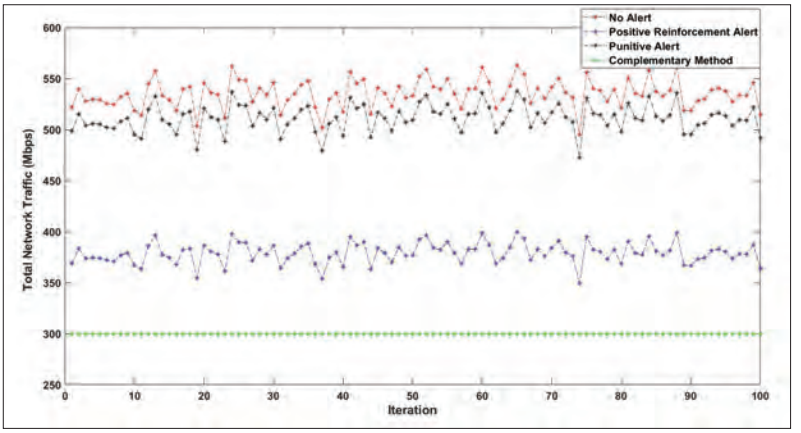


FIGURE 4. Cellular traffic variations over time for a given load.

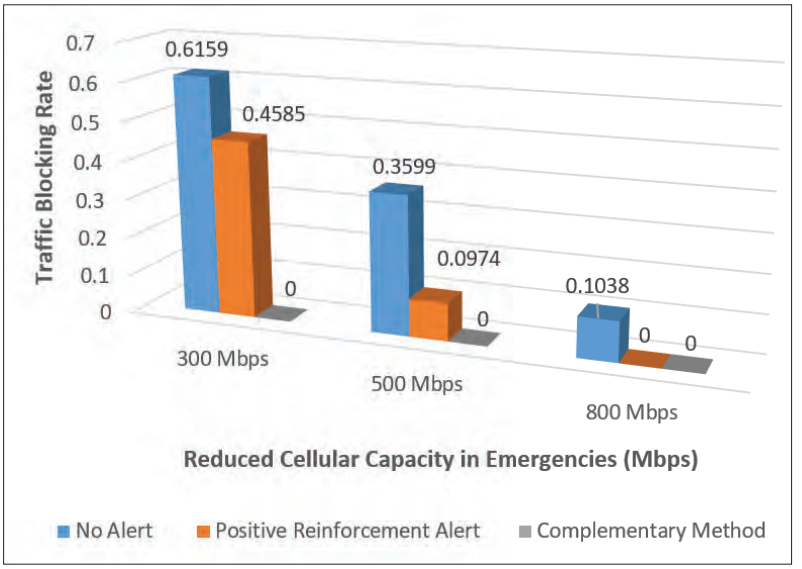


FIGURE 5. Traffic blocking rate.

Acknowledgments

This work is supported in part by the NSF under Grant No. ACI1541069. The authors thank Arnold Glass and Margaret Ingate for their insights that improved the design of the survey.

References

[1] T. Frank, "Cellphone Service Must Be Restored Quicker after Hurricanes," *E & E News*; https://www.scientificamerican.com/article/cell-phoneservice-must-be-restored-quicker-after-hurricanes/, accessed Mar. 14, 2022.

[2] E. Wax-Thibodeaux, "'Cajun Navy' Races from Louisiana to Texas, Using Boats to Pay It Forward," *Washington Post*, 2017; https://www.washingtonpost.com/national/cajun-navy-races-from-louisiana-to-texas-using-boats-topay-it-forward/2017/08/28/1a010c8a-8c1f-11e7-84c002cc069f2c37_story.html, accessed: Mar. 14, 2022.

[3] "Integrated Public Alert & Warning System," FEMA; https://www.fema.gov/emergencymanagers/practitioners/integrat-

ed-public-alert-warningsystem, accessed Mar. 14, 2022.
[4] FCC, "Wireless Emergency Alerts (WEA)"; https://www.fema.gov/emergency-managers/practitioners/integrated-public-alert-warning-system/public/wirelessemergency-alerts, accessed Mar. 14,2022.
[5] M. A. Youssef *et al.*, "Message Content Affects the Willingness to Comply with Voluntary Restriction of Resource Use," poster presented at easternpsychological.org, 2016 summit; https://www.easternpsychological.org/i4a/pages/index.cfm? pageid=3550, 2016, accessed Mar. 14, 2022.
[6] M. Yousefvand, *User Centric and Network Centric Approaches for Resource and Emergency Alert Optimization in Wireless Networks*, Ph.D. dissertation, ECE Dept., Rutgers Univ., 2021.
[7] 3GPP TS 38.300, "NR; NR and NG-RAN Overall Description", 3GPP Tech. Spec. Group Radio Access Network, v. 16.8.0, 2021; https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3191, accessed Mar. 27, 2022.
[8] M. A. Casteel and J. R. Downing, "Assessing Risk Following a Wireless Emergency Alert: Are 90 Characters Enough?," *J. Homeland Security and Emergency Management*, vol. 13, no. 1, 2016, pp. 95–112.
[9] FCC. "16-127 Document on WEAs," Sept. 2016; https://docs.fcc.gov/public/ attachments/FCC-16-127A1.pdf.
[10] S. Barclay, "Delivering Targeted Alerts Advancing the Wireless Emergency Alerts (WEA) 3.0 System," 2019, PowerPoint slides; https://www.atis.org/wpcontent/uploads/01_news_events/webinarpptslides/ATIS_WEA3.0webinar.pdf.
[11] Verizon Household Data Usage Calculator; https://www.verizonwireless.com/freedom/datacalculator.html, (2020), accessed Apr. 16, 2020.
[12] D. Lambropoulos, M. Yousefvand, and N. Mandayam, "Tale of Seven Alerts: Enhancing Wireless Emergency Alerts (WEAs) to Reduce Cellular Network Usage During Disasters," arXiv:2102.00589 [cs.HC], 2021.
[13] Y. Qu *et al.*, "Online Community Response to Major Disaster: A Study of Tianya Forum in the 2008 Sichuan Earthquake In 2009," *42nd Hawaii Int'l. Conf. System Sciences*, 2009.
[14] I. Shklovski, L. Palen, and J. Sutton, "Finding Community Through Information and Communication Technology in Disaster Response," *Proc. 2008 ACM Conf. Computer Supported Cooperative Work*, 2008, pp. 127–36.
[15] F. Xu *et al.*, "Understanding Mobile Traffic Patterns of Large Scale Cellular Towers in Urban Environment," *IEEE/ACM Trans. Net.*, vol. 25, no. 2, 2015, pp. 1147–61.

Biographies

Mohammad Yousefvand (my342@winlab.rutgers.edu) is a senior systems engineer at Qualcomm in San Diego, California. He holds Ph.D. and Master's degrees in electrical engineering from Rutgers University, and a Master's degree in computer engineering from Amirkabir University of Technology (Tehran Polytechnic). His research interests include cross-layer optimization of 5G networks, user scheduling, resource allocation, machine learning, game theory, and prospect theory.

Demetrios Lambropolous (dpl60@scarletmail.rutgers.edu) is a Ph.D. candidate in computer engineering at the Electrical and Computer Engineering Department, Rutgers University, New Brunswick, New Jersey. He is currently working on several projects under the supervision of Prof. Narayan B. Mandayam at the Wireless Information Network Laboratory, Rutgers University. His research interests include machine learning, resource allocation, human-computer interaction, and security.

Narayan B. Mandayam [F] (narayan@winlab.rutgers.edu) is a Distinguished Professor and Chair of Electrical and Computer Engineering at Rutgers University, where he also serves as Associate Director of WINLAB. His recent interests include enabling privacy in IoT, building resilience in smart city infrastructures, and trustworthy knowledge creation on the Internet. He received the 2015 IEEE ComSoc Advances in Communications Award, the 2014 IEEE Donald G. Fink Award, and the 2009 Fred W. Ellersick Prize from the IEEE Communications Society. He is also a recipient of the Peter D. Cherasia Faculty Scholar Award from Rutgers University (2010), the National Science Foundation CAREER Award (1998), the Institute Silver Medal from the Indian Institute of Technology (1989), Kharagpur, and its Distinguished Alumnus Award (2018). He is a Distinguished Lecturer of the IEEE.

IEEE World Forum on Internet of Things

The 8th IEEE World Forum on the Internet of Things (IoT)

26 October–11 November 2022

Yokohama, Japan

Pacifico Yokohama Convention Center

Hybrid: In-Person and Virtual

Sustainability and the Internet of Things

Sponsored by the Multi-Society IEEE IoT Initiative

The 8th IEEE World Forum on the Internet of Things (WF-IoT 2022) will be held 26 October–11 November 2022 in Yokohama, Japan and virtually. The content of the program will be aimed at growth of IoT across the world and the reduction of harmful and environmental impacts.

WF-IoT 2022 will showcase the latest from the research and academic community through a broad program of papers and presentations on the latest technology developments and innovations in the many fields and disciplines that drive the utility and vitality of IoT solutions and applications.

The program will also feature strong involvement from the public sector and industry aimed at deepening the understanding, the necessary dialog, and actions needed to accelerate the adoption and deployment of IoT.

IoT Vertical Sessions:

- Agriculture
- Energy and Power
- Industry and Manufacturing
- Smart Cities
- Transportation

IoT Topical Area Sessions:

- Artificial Intelligence
- Communications and Networking
- Computing and Processing
- Robotics
- Sensors and Sensor Systems

Don't miss this exciting educational and networking opportunity to join hundreds of IoT experts and enthusiasts as they present research results, share visions and ideas, and deliver updates on latest IoT technologies.

Visit wfiot2022.iot.ieee.org for more information and to register.

Fronthaul Compression Control for Shared Fronthaul Access Networks

Sandra Lagén, Xavier Gelabert, Andreas Hansson, Manuel Requena, and Lorenza Giupponi

The authors focus on FH compression control strategies for multiple-cell/multiple-user scenarios sharing a common FH link. They propose various methods for sounding reference signal handling and analyze different FH-aware modulation data compression and scheduling strategies.

Abstract

There is a widely held belief that future radio access network architectures will be characterized by increased levels of virtualization, whereby base station functionalities, traditionally residing at a single location, will be scattered across different logical entities while being interfaced via high-speed fronthaul (FH) links. For the deployment of such FH links, operators are faced with the challenge of maintaining acceptable radio access performance while at the same time keeping deployment costs low. A common practice is to exploit statistical multiplexing by allowing several cells to utilize the same FH link. As a result, in order to cope with the resulting aggregated traffic, different techniques can be used to reduce the required FH data rates. Herein, we focus on FH compression control strategies for multiple-cell/multiple-user scenarios sharing a common FH link. We propose various methods for sounding reference signal (SRS) handling, and analyze different FH-aware modulation data compression and scheduling strategies. Considering a full system setup, including the radio and FH access networks, numerical evaluation is conducted using a 5G NR system-level simulator implemented in ns-3. Simulation results show that under stringent FH capacity constraints, optimized modulation compression strategies provide significant user-perceived throughput gains over baseline strategies (between 5.2× and 6.9×). On top of them, SRS handling methods achieve additional 2 to 41 percent gains.

Introduction

The design of future radio access networks (RANs) often needs to fulfill the demanding requirements across different and competing axes. While increasing the spectral efficiency has a major impact on the perceived quality of service for the end user, implementing a scalable and low-power solution is often regarded as important for network operators. Furthermore, efficient use of computing resources via pooling and the ability to provide cross-layer solutions are also desired features. With this in mind, the centralized RAN (C-RAN) architecture paradigm [1] has emerged and is being considered by the 3rd Generation Partnership Project (3GPP) and Open-RAN (O-RAN) as a key design alternative in next-generation RANs. Among its features, C-RAN advocates for the dissagregation of baseband processing functions between different physical entities that can be either distributed and residing close to the antenna or centralized in a given location. Specifically, a base station, also known as a gNB in 3GPP 5G New Radio (NR), will break up into a centralized unit (CU) communicating with at least a distributed unit (DU) via the so-called midhaul interface [2]. In turn, several radio units (RUs) will interface toward a DU via the fronthaul (FH) [3]. When designing and deploying C-RAN, it is important to consider both capacity constraints and latency requirements of the FH [4]; more so considering the increased bandwidths in 5G NR, in addition to antenna densification, increased modulation orders, and enhanced carrier aggregation features [5]. All of these contribute to the increase in the required FH capacity [6].

In general, the dimensioning of the FH may respond to the peak rate requirements given during the planning phase of a specific network technology deployment (e.g., 4G LTE). Nonetheless, under normal network operation, several reasons cause the FH to undergo resource underprovisioning at specific moments in time [7]. One example is the ever growing adoption of new features, new functional splits, or new algorithms requiring increased information exchange between RUs and DUs/CUs. Other examples are to allow a seamless rollout of new radio access technologies (e.g., 5G NR) or to facilitate a continuous layout of low-power small cells in specific traffic-demanding areas while at the same time allowing this new data to be handled through the available and pre-dimensioned FH network. From the equipment vendor's viewpoint, some interest may be rooted in offering a set of new features requiring minimal disturbance on the existing FH network, where only software updates would be necessary to upgrade firmware, algorithms, and so on. On the contrary, a dimensioning change in the FH network may be seen as a costly measure, that is, by exchanging optical interface adapters, network switches, and, at worst, optical fiber layouts themselves (e.g., switching from single-mode to multi-mode). All the above provides a good motivation to consider the case where the FH can run into a capacity underprovisioning problem.

To lessen the demand in FH capacity and address the above-mentioned FH underprovisioning problem, FH compression schemes become essential [1]. Briefly, FH compression involves the partial reduction, or total removal, of information sent over the FH. FH compression methods

Digital Object Identifier: 10.1109/MCOM.001.2100959

Sandra Lagén, Manuel Requena, Lorenza Giupponi are with Centre Tecnològic de Telecomunicacions de Catalunya (CTTC/CERCA), Spain; Xavier Gelabert, Andreas Hansson are with Huawei Technologies Sweden AB, Sweden.

0163-6804/22/$25.00 © 2022 IEEE

have received wide attention from both the information theory and signal processing communities, in particular after the emergence of C-RAN architectures in 4G LTE. Recently, more practical schemes have been defined in 3GPP 5G NR and O-RAN [3]. Among the envisioned techniques, *modulation compression* is highlighted given its ability to reduce the required FH capacity due to constraining the modulation order. This can be effectively achieved with no degradation of the transmitted samples sent over the FH network and with reduced algorithmic complexity [6].

Besides FH compression, aimed at jointly reducing the deployment and operational costs along with fostering deployment scalability, the ability to multiplex and aggregate data from multiple cells over a single shared FH interface is of great interest to mobile network operators [8]. In this case, the same fixed FH link is shared by multiple cells (i.e., data from/towars several RUs are multiplexed by using a layer-2 switch over a single FH link toward/from their respective DUs), as shown in Fig. 1. These scenarios present multiple technical challenges arising from the shared FH use by data originating or terminating from/to different cells. Essentially, effective FH compression control schemes have to especially consider and exploit the fact that many cells share a fixed capacity full duplex FH link, and should enhance the multiplexing gain by using clever combinations of different compression techniques, providing efficient methods to control the use of FH resources while minimizing air interface performance degradation. Consequently, an evaluation approach relying on end-to-end system-level simulations in a multi-cell environment is carried out in this work. In [9], baseline along with improved FH-aware packet scheduling methods of dynamic modulation compression were derived. Therein, both the scheduling and modulation compression decisions were dynamically adjusted according to the monitored FH capacity.

When considering the full system, even if we focus on downlink data transmission, the main part of the data bulk sent through the FH interface comes from the physical downlink shared channel (PDSCH) (used to send data) and the uplink sounding reference signals (SRSs) (used to estimate the channel). Modulation compression helps reduce the PDSCH bulk part in the FH downlink link. However, when using SRS-based channel estimation for beamforming/precoding design, as standardized for time-division duplex (TDD) 5G systems, the FH uplink utilization can be reduced by selectively compressing/removing unnecessary SRSs, which may further impact the beamforming/precoding design in the downlink. Here, *SRS handling* techniques can alleviate the FH uplink load by properly handling the allocation of SRS signals through the full duplex FH interface. In this article, different from [6] where disaggregated architectures and FH compression methods were reviewed, and from [9] where dynamic modulation compression methods were proposed to reduce only the PDSCH bulk, we provide a summary and a new vision of integral solutions for dynamic FH compression control in shared FH architectures. In particular, we review methods for FH-aware downlink data scheduling, and we provide a new study for SRS handling methods in the uplink. Finally, the numerical evaluation of the aforementioned schemes is carried out using a seasoned dynamic system-level simulator developed in ns-3 [10].

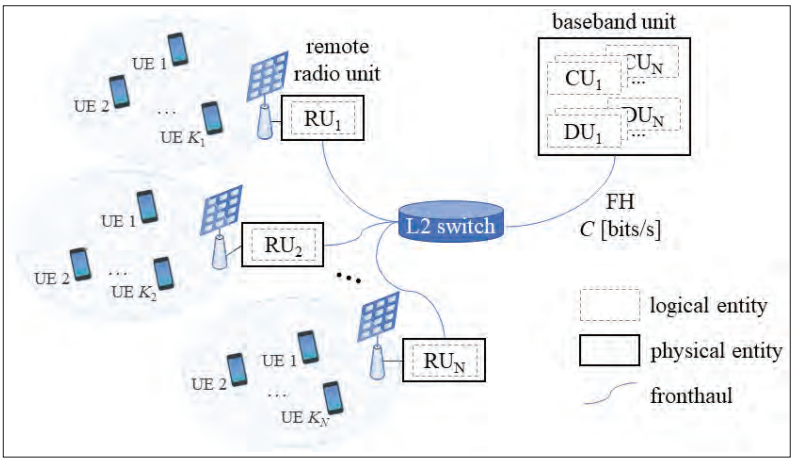


FIGURE 1. Deployment scenario. Multiple cells and multiple UEs per cell. Hybrid C-RAN architecture with multiple CUs, DUs, and RUs. Star FH topology with a shared full duplex FH interface of limited capacity in downlink and uplink.

The rest of the article is structured as detailed hereafter. We discuss the system model and overview FH compression methods, with special attention to SRS handling methods for uplink SRSs. We describe the simulation scenario and assess the end-to-end performance. Finally, we highlight future research lines and conclude the article.

Fronthaul Compression Control

This section introduces the system model and discusses solutions for shared FH architectures integrating FH-aware scheduling methods and SRS handling methods, to reduce the FH downlink and FH uplink loads, respectively.

System Model

We consider herein a multiple-cell/multiple-user-equipment (UE) cellular deployment following a C-RAN architecture and consisting of co-located CUs/DUs and geographically distributed RUs. We consider the PDCP-RLC split (a.k.a. Option 2) for the CU-DU [2] and the intra-PHY split (Option 7.2x) for the DU-RU [3], as per 3GPP and O-RAN specifications, respectively. Regarding the CU/DU/RU deployment, we follow the so-called Scenario B, as highlighted by O-RAN [11]. In this case, CUs and DUs (for all cells) are located together in a centralized location (edge or regional cloud), whereas RUs are scattered at operator-owned cell sites. Consequently, the centralized entity implements the high-PHY, medium access control (MAC), and above processing for all cells, while the RU of each site executes the low-PHY and RF processing of every cell [12]. DUs are interconnected with the RUs through a low-layer full-duplex FH interface of limited capacity in downlink and uplink directions, according to a star FH topology [3]. This way, RUs share the same full duplex FH link. The deployment scenario is illustrated in Fig. 1.

We focus on FH compression control for downlink data transmission in multi-cell TDD systems with a shared full duplex FH interface. In particular, we:

1. Analyze FH-aware scheduling methods

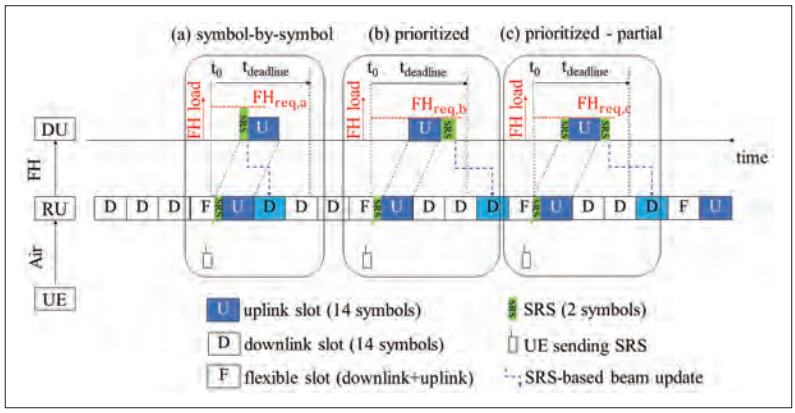


FIGURE 2. SRS transmission through the full-duplex FH interface, when using a TDD pattern of [D D D F U] in the air interface (i.e., three downlink slots, followed by a flexible slot and one uplink slot) and SRSs being sent in the F slots.: a) symbol-by-symbol transmission; b) prioritized FH transmission; c) prioritized FH transmission with partial SRS transmissions.

based on modulation compression to meet the available FH downlink capacity
2. Propose SRS handling methods to meet the available FH uplink capacity

Downlink FH-Aware Scheduling Methods for PDSCH

In typical cellular systems, the downlink modulation and coding scheme (MCS) for each UE is determined based on the reported channel quality indicator (CQI). Given the MCSs and the amount of data in the RLC buffers, the MAC scheduler decides the number of resource blocks (RBs) allocated to each UE by following specific scheduling rules (e.g., proportional fair). To meet the shared FH downlink capacity constraint, several FH-aware scheduling methods have been proposed in the literature. Two baseline options are:

1. Dropping packets at the PHY layer
2. Postponing scheduling decisions at the MAC layer [9]

Another option, is to limit the MCS per cell [6] using modulation compression. To enhance these solutions, centralized optimized methods were proposed in [9], in which the resource allocation (i.e., number of RBs) and modulation compression (i.e., MCS) of each UE are dynamically set. These methods are reviewed in what follows.

Drop Packets at High-PHY: Assuming typical MCS selection and RB assignment at the MAC scheduler, packet dropping can be implemented at the high-PHY layer in the DUs. In this case, a centralized logic decides to drop those MAC packet data units (PDUs) (including new data and/or hybrid automatic repeat request, HARQ, retransmissions and their associated control across all cells) that cannot fit in the available shared FH downlink capacity [9].

Postpone Scheduling Decisions at MAC: Assuming typical MCS selection and RB assignment at the MAC scheduler, discarding/dropping MAC scheduling decisions can be executed at the MAC layer in the DUs. This way, data is not dropped, but its transmission (or retransmission) is postponed. In this case, a centralized logic decides to drop/discard those scheduling decisions (related to new data and/or HARQ retransmissions across all cells) that cannot fit in the available shared FH downlink capacity, for which its associated transmission is postponed [9].

MCS Limits at MAC: By exploiting semi-static modulation compression [6], the system can limit the maximum MCS (and thus the maximum modulation order) that is allowed per cell according to the available shared FH downlink capacity. Note that per-cell MCS limits can be combined with dynamic methods, like drop and postpone strategies. In particular, their joint operation may result in fewer packet drops and scheduling decision postponements because of the inherent FH load reduction when using lower MCSs.

Resource Allocation and MCS Optimization at MAC: By using dynamic modulation compression, a centralized control entity can manage the MAC schedulers of all the cells (placed in the collocated DUs in Fig. 1) and determine the most appropriate MAC scheduling decisions (including scheduling of users, MCS assignment, and resource allocation) across all cells dynamically so that the shared FH downlink capacity is properly exploited and certain QoS per user is satisfied. In particular, two solutions were derived in [9] to optimize the RB allocation and the modulation compression applied to each UE of each cell for every time instant.

Uplink FH-Aware Handling Methods for SRSs

In TDD systems, SRSs can be used for beam management in downlink and uplink, due to beam and channel reciprocity. In particular, SRS receptions at the base station are typically used to estimate the channel and then determine the beamforming/precoding. To get an accurate acquisition of the SRS signal, the bulk needed to send SRSs through the FH uplink interface (RU-to-DU) can be very large, since its size depends on the number of antennas used for channel estimation and the number of resource elements carrying SRS samples.

A key observation is that SRS signals, different from downlink/uplink data in PDSCH/PUSCH, are non-delay-sensitive and do not need to be transmitted through the FH as soon as they arrive at the RU. Basically, we can exploit the fact that the FH interface is full duplex, while the air interface is half duplex; thus, by leveraging on the TDD pattern, SRSs can be sent through the FH uplink interface when there is a downlink slot.

Figure 2 illustrates examples of the impact on the peak FH bandwidth requirements of SRS transmissions, assuming that for a given SRS signal, a certain amount of bits needs to be processed at the DU before some target deadline. There are two main examples of how to meet the latency constraint:

- Symbol-by-symbol FH transmission (Fig. 2a): After each symbol is received at the RU, it is immediately transmitted over the FH. The required FH capacity is then given by the maximum between the SRS and PUSCH bulk requirements. This option causes high load peaks on the FH, and makes it so that the dimensioning of the required FH capacity is determined by SRS bulks, which require more samples than PUSCH, especially for multi-antenna and wide-bandwidth systems.
- Prioritized FH transmission (delayed transmission) (Figs. 2b and 2c): A certain delay is allowed when conveying SRSs over the FH. For example, in Fig. 2b, PUSCH transmission is prioritized, and SRS samples are buffered

at the RU and transmitted later over a time that may be longer than 2 symbols but still allows the SRSs to be processed before its deadline. A further enhancement is to partially transmit the SRSs before and after the full PUSCH bulk, as shown in Fig. 2c. In both cases, we need buffering at the RU for SRSs, and the FH uplink capacity requirement is determined by the PUSCH bulk, not by the more demanding SRS bulk.

Prioritized FH transmissions reduce the FH uplink capacity requirement at the cost of a different delay on the reception of the full SRS signal at the DU. This consequently implies a delay on the beamforming/precoding update (illustrated by the light blue downlink slot in Fig. 2), which may affect the downlink end-to-end performance and should be properly evaluated through system-level simulations.

In the case of multiple RUs sharing the same FH interface, and potentially overlapping SRS transmissions (e.g., because different cells/RUs use the same TDD pattern), SRS handling methods need to be designed. Two options appear: time multiplexing or frequency multiplexing. In the frequency multiplexing option, all the SRS bulks experience the same FH delay, as shown in Figs. 3a and 3b. We can use a worst case partition of the available FH bandwidth among the multiple RUs that share the FH (as shown in Fig. 3a, for the case of 3 RUs sharing the FH interface, where only 2 RUs send SRSs at the same time). In this case, the FH capacity may not be fully exploited in the uplink direction. Otherwise, we can adopt a dynamic FH bandwidth allocation to the RUs that have data to be sent on a particular slot (as shown in Fig. 3b), where the total delay can be reduced compared to the hard bandwidth distribution option. On the other hand, through the time multiplexing option, dynamicity is naturally achieved. In this case, SRS bulks are sent sequentially (Fig. 3c) so that some of them are received more quickly at the DU/CU for processing. This option allows the beam update to be done sooner compared to the frequency multiplexing options. Here, SRS priority handling methods are needed to decide the order/priority to send the UEs' SRS bulks.

End-to-End Simulation

For the evaluation, the ns-3 5G-LENA system-level simulator is used [10]. We extended the 5G-LENA simulator with a new centralized intelligence that controls the MAC/PHY operations of all the DUs and implements the proposed FH-aware scheduling procedures and SRS handling methods.

Scenario

We consider a hexagonal site deployment with three sites, according to an Urban Micro scenario. Each site is composed of 3 cells and 3 uniform planar antenna arrays, covering 120° in azimuth each. Frequency reuse 1 is assumed. The rest of the deployment and network parameters are:

- Number of cells (RUs): 9
- Number of UEs per cell: 10
- Inter-site distance: 200 m
- RU antenna height: 10 m
- RU transmit power: 30 dBm
- RU antenna: 5 × 2 directional elements
- UE antenna height: 1.5 m
- UE antenna: 1 isotropic element
- Carrier frequency: 2 GHz

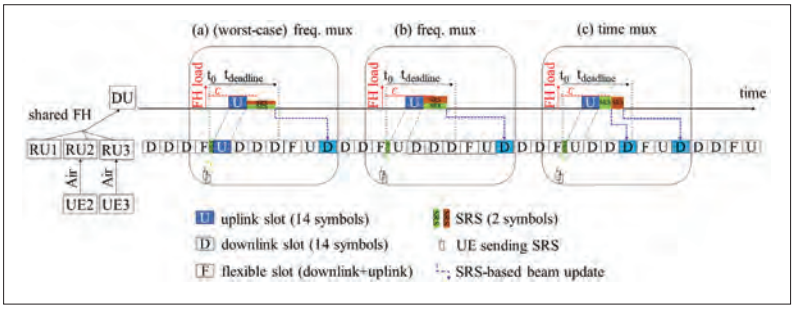


FIGURE 3. SRS handling methods for SRS bulk transmission through the FH. As an example, the FH is shared among three RUs, two being active at the F slot: a) (worst case) fixed frequency multiplexing; b) dynamic frequency multiplexing; c) dynamic time multiplexing.

- Bandwidth: 100 MHz
- Numerology: 1 (30 kHz subcarrier spacing)
- RB overhead: 0.04
- Duplexing mode: TDD, with pattern [D D D F U].
- SRS: in F slot, spanning over 1 orthogonal frequency-division multiplex (OFDM) symbol. Two SRS periodicities:
  - 50 ms (*SRS config1*)
  - 25 ms (*SRS config2*)
- MAC scheduler: Round-Robin
- MCS Table: 2 (up to 256-QAM: quadrature amplitude modulation)
- Channel update period: 40 ms
- HARQ: incremental redundancy, 20 HARQ processes
- RLC: unacknowledged mode
- Transport protocol: UDP
- Traffic: File Transport Protocol (FTP) Model 1 [1, Sec. A.2.1.3.1], YouTube video characterization [14]:
  - File size: 50 kB
  - File generation rate: 50 files/s
- FH: start topology, shared full-duplex FH link of 0.5 Gb/s capacity in each direction (downlink/uplink)[1]
- Simulation duration: 10 s

As the key performance indicator we consider the user-perceived throughput (UPT), measured at the IP layer. The UPT corresponds to the fraction between the received bytes per file and the time period needed to complete the file transfer.

Results

In the end-to-end evaluation, we assess the impact of using different FH compression control methods. For downlink data, we consider the following FH-aware scheduling methods (described earlier):

- Drop: High-PHY drop of MAC PDUs
- Postpone: Discard MAC scheduling decisions
- RB: RB assignment optimization per active UE
- MCS: MCS optimization per active UE

We evaluate each strategy in combination with the following SRS handling methods (detailed earlier):

- fixedDelay: The FH uplink bandwidth is equally distributed among all the cells that share the FH.
- dynDelay freqMux: The FH uplink bandwidth is equally distributed among active cells (delay is time-dependent).
- dynDelay timeMux: The FH uplink bandwidth is fully used by each SRS bulk (delay is time- and cell-dependent).

Figure 4 shows the cumulative density function (CDF) of the UPT (in megabits per second) when using SRS config, for different scheduling strate-

[1] The FH capacity has been derived based on the peak FH throughput computation. Considering 9 cells, all RBs/symbols used with maximum MCS, and multiplexing gain of 0.5, the total peak FH throughput results 3.6 Gb/s. Thus, we select 0.5 Gb/s value, to increase the probability that the system is constrained by the FH.

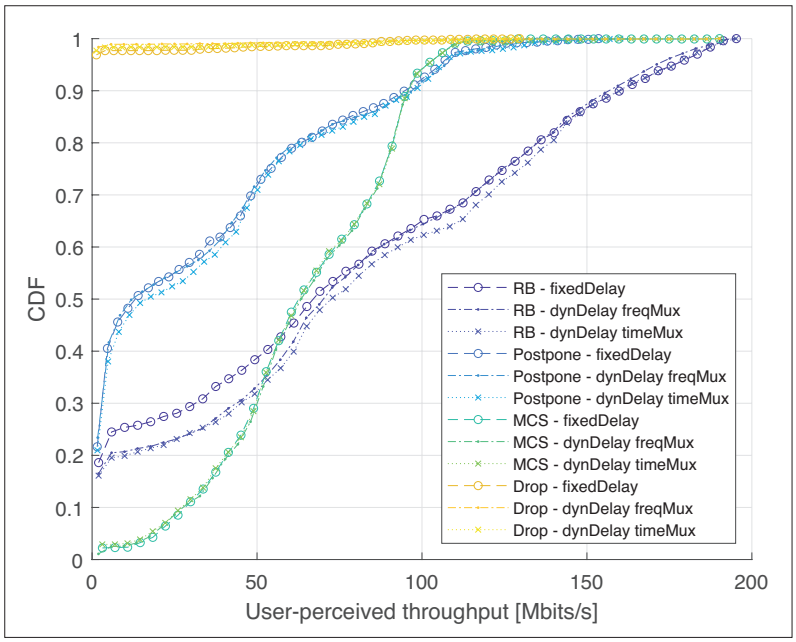


FIGURE 4. UPT CDF (Mb/s), for different scheduling and SRS handling methods. SRS config1.

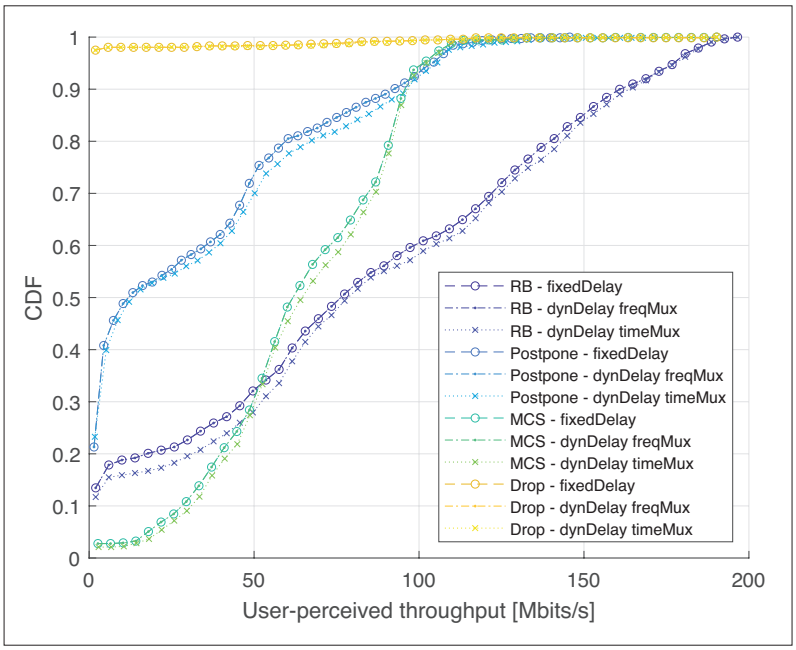


FIGURE 5. UPT CDF (Mb/s), for different scheduling and SRS handling methods. SRS config2.

gies and SRS handling methods. Figure 5 displays the same for SRS config2.

Regarding FH-aware scheduling methods, the Drop technique exhibits very bad end-to-end performance. This is because it performs the scheduling as usual, and in many cases, the scheduled data allocations cannot fit in the available FH bandwidth, implying frequent data drops. The performance is improved with the Postpone strategy, which postpones the scheduled allocations for moments when FH capacity is available, leading occasionally to increased delay. In the considered scenario, characterized by a tight FH capacity, a clear advantage is observed with RB and MCS optimization strategies over baseline strategies (Drop and Postpone) (Figs. 4 and 5). The benefits are significant in all percentiles of the UPT, ranging from 5.2× to 6.9× in the mean UPT over the Postpone strategy. Indeed, the MCS strategy outperforms the RB strategy in the 5th percentile UPT because, by reducing the modulation order, higher robustness is achieved against propagation/interference variations. Conversely, RB strategy outperforms MCS in the 95th percenttile UPT, because it provides a more efficient RB distribution due to a fine-grained control mechanism, thus enabling a larger amount of data to be served.

Regarding the SRS handling methods, we observe that the impact of using beams which are not well adjusted to the channel is appreciable in the RB optimization strategy, while for other strategies the impact is reduced (Figs. 4 and 5). This is because Drop/Postpone strategies are dropping/postponing many packets, so the performance is dominated by the losses; while the MCS strategy is more robust to signal-to-interference-plus-noise ratio (SINR) degradation.

The impact of the different SRS handling methods on the UPT depends on the SRS configuration. Specifically, in SRS config1 (Fig. 4), both the frequency and time multiplexing with dynamic adaptation improve the UPT performance of the worst case frequency multiplexing, and both provide similar performance. This is because with the considered frame pattern and SRS periodicity (50 ms), there are 20 available slots for SRSs within the SRS periodicity, and the deployment considers 10 UEs per cell. Therefore, not all the cells are active at each SRS opportunity. In this way, the dynamic freqMux also allows reducing beam update delays compared to the worst case freqMux option, and gets similar UPT performance as the timeMux option, which has lower delays, because SRS bulks are sent one after the other. Under SRS config2, we have considered the same scenario but with an SRS periodicity of 25 ms, which results in 10 opportunities for SRSs. All SRS opportunities are then used by one of the UEs to send SRSs. In this case, as expected and shown in Fig. 5, the two frequency multiplexing options (with fixed or time-dependent delays) achieve the same end-to-end performance, because the two options are equivalent. Interestingly, the time multiplexing option outperforms both frequency multiplexing options, since a major part of the SRS bulks experience lower delay updates. In summary, dynamic SRS handling methods achieve mean UPT gains ranging from 2 to 41 percent.

Finally, the presented results have allowed interesting observations regarding the behaviors of FH-scheduling and SRS-handling methods as a function of the type of served UEs (i.e., cell edge, cell middle, or cell center). A summary of main conclusions is shown in Table 1. For cell-edge UEs, the best option is to use MCS strategy, and the SRS handling method does not impact the performance. For cell middle and cell center UEs, the best option is to use RB strategy in downlink. In uplink, the cell center UEs are not affected by the SRS handling method, while for cell middle UEs, the recommended option is the time multiplexing.

Future Research Directions

Based on the presented study, analysis, and

obtained simulation results, we envision the following research lines:

- SRS priority handling methods: The results from earlier show that users in different conditions (cell edge/middle/center) are affected differently by the delay updates of the beams. Accordingly, if a time-multiplexing option is adopted, clever SRS priority handling methods for the SRS bulks have to be defined. Under shared FH capacity, SRS bulks associated with specific UEs could be prioritized to reduce their delays. For example, when using an RB strategy, cell edge UEs could be prioritized, because they are more affected by a delay increase in the beam update. Instead, when using a combined MCS/RB strategy, cell middle UEs could be prioritized. A control entity at the FH interface could implement the SRS priority handling method by knowing the SINR associated wotj each SRS bulk to distinguish among cell center/middle/edge UEs.
- SRS control methods: The control entity could also keep track of the actual FH delay for each SRS bulk, leading to SRS control methods. For example, if the buffering delay surpasses the channel coherence time, such an SRS bulk could be dropped from the buffer of packets to be sent through the FH, because when the DU receives it, the measurement will already be outdated. This would leave FH capacity available for other transmissions.
- SRS priority handling methods with partial transmissions and partial beam updates: Partial transmissions of the SRS blocks, as shown in Fig. 2c, constitute an improvement for shared FH interfaces. In this case, part of the SRS bulks of different UEs can be sent earlier so that the DU/CU can do a first channel estimation and beamforming update with partial information (e.g., half or part of the SRS samples). Later, once the full SRS information is sent through the FH, a second beamforming update can be implemented based on complete information.
- Uplink data FH compression: In the present study, we have focused on compression of FH information related to downlink data transmissions, for which the downlink data (in downlink) and SRS (in uplink) constitute the bulk FH part. Future studies could include compression of uplink data in PUSCH.
- Joint flexible splits and FH compression: There has been wide interest in flexible functional split selection recently. However, the interaction between the split selection and the scheduling/resource allocation strategies has been less studied. An interesting area for further research is to analyze joint strategies that optimize the functional split and the FH compression control for shared FH multi-cell scenarios.

Conclusions

In this article, we have presented an integral design and a thorough end-to-end evaluation of shared FH scenarios where multiple FH compression control techniques are proposed. In particular:

1. We have analyzed FH-aware scheduling methods to compress user data that goes through the downlink FH.
2. We have proposed SRS bulk handling methods to handle uplink SRS bulks that go through the uplink FH.

	Cell edge	Cell middle	Cell center
FH-aware scheduling	MCS	RB	RB
SRS handling	All	dynDelay timeMux	All

TABLE 1. Best FH-aware scheduling strategy and best SRS handling method, depending on the UE position within a cell.

Then end-to-end simulations over a 5G-aligned scenario have been presented. In multi-cell scenarios with shared FH link, we have evaluated the impact of four main FH-aware scheduling methods for downlink data compression — Drop, Postpone, RB, and MCS — combined with three methods for SRS handling: fixed frequency multiplexing, dynamic frequency multiplexing, and dynamic time multiplexing. Results have shown that when there is tight FH capacity, centralized and optimized scheduling strategies (MCS and RB methods) are essential to maintain an acceptable end-to-end user experience. SRS handling methods are shown to affect the RB optimization strategy, for which our results have exhibited that the time-multiplexing option always provides the best performance and improves all the other SRS handling methods for configurations in which all the SRS opportunities are used to send SRSs. However, when not all the SRS opportunities are used to send SRSs, dynamic frequency multiplexing can also achieve similar performance. Interestingly, the degradation in the end-to-end performance depends on the quality/condition of the target UE, and it is more pronounced in the cell edge/middle users, which get a lower SINR in the downlink as a result of the path loss degradation and larger errors in the SRS-based channel estimation, for which future research lines have been highlighted.

Based on our findings above, we advocate for the following recommendations. First, operators and vendors should seriously consider the FH underprovisioning problem, whereby a properly dimensioned FH at the planning stage may become underprovisioned over time. Second, considering shared FH link segments is key to exploit multiplexing gains arising from traffic inhomogeneities. Third, when capacity-limited FH problems arise, leveraging FH compression strategies helps alleviating the problem. We conclude that dynamic compression of data is essential to maintain acceptable user experience, while at the same time noting that compression of reference signals (especially SRSs) plays a relevant role.

Acknowledgments

This work has been partially funded by Huawei Technologies and Spanish MINECO grant TSI-063000-2021-56/57 (6G-BLUR).

References

[1] M. Peng *et al.*, "Recent Advances in Cloud Radio Access Networks: System Architectures, Key Techniques, and Open Issues," *IEEE Commun. Surveys & Tutorials*, vol. 18, no. 3, 2016, pp. 2282–2308.
[2] 3GPP TS 38.401, TSG RAN, "NG-RAN; Architecture Description," Release 16, v. 16.1.0, Mar. 2020.
[3] O-RAN Fronthaul Working Group, "Technical Specification; Control, User and Synchronization Plane Specification)," v. 03.00, Apr. 2020.

[4] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks," *IEEE Commun. Surveys & Tutorials*, vol. 21, no. 1, 2019, pp. 146–72.
[5] S. Parkvall *et al.*, "NR: The New 5G Radio Access Technology," *IEEE Commun. Standards Mag.*, vol. 1, Dec. 2017, pp. 24–30.
[6] S. Lagen *et al.*, "Modulation Compression in Next Generation RAN: Air Interface and Fronthaul Trade-offs," *IEEE Commun. Mag.*, vol. 59, no. 1, Jan. 2021.
[7] M. Peng *et al.*, "Fronthaul-Constrained Cloud Radio Access Networks: Insights and Challenges," *IEEE Wireless Commun.*, vol. 22, no. 2, Apr. 2015, pp. 152–60.
[8] A. Umesh *et al.*, "Overview of O-RAN Fronthaul Specifications," Special Articles on Standardization Trends towards Open and Intelligent Radio Access Networks, *NTT Docomo Tech. J.*, vol. 21, July 2019.
[9] S. Lagen *et al.*, "Fronthaul-Aware Scheduling Strategies for Dynamic Modulation Compression in Next Generation RANs," *IEEE Trans. Mobile Computing*, Nov. 2021.
[10] 5G-LENA Project; https://5g-lena.cttc.es/.
[11] O-RAN Alliance White Paper, "O-RAN Use Cases and Deployment Scenarios," Feb. 2020.
[12] 3GPP TR 38.801, TSG RAN, "Study on New Radio Access Technology: Radio Access Architecture and Interfaces," Release 14, v. 14.0.0, Mar. 2017.
[13] 3GPP TR 36.814, "Evolved Universal Terrestrial Radio Access (EUTRA); Further Advancements for E-UTRA Physical Layer Aspects," Release 9, v. 9.2.0, Feb. 2017.
[14] Samsung, "Application of Stochastic Geometry in Modeling Future LTE-A and 5G Wireless Networks"; https://web.ma.utexas.edu/conferences/simons2015/Modeling3GPP Samsung Final.pdf.

Biographies

Sandra Lagén (sandra.lagen@cttc.es) holds a Ph.D. from UPC (2016). She is a senior researcher and head of the Open Simulations research unit at CTTC.

Xavier Gelabert (xavier.gelabert@huawei.com) is a senior research engineer at Huawei Technologies Sweden AB. He has 15+ years of experience working across RAN L1, L2, and L3, as well as 3GPP standardization.

Andreas Hansson (andreas.hansson@huawei.com) is a principal baseband software engineer within Huawei Technologies Sweden AB, working with software architecture, modeling, and systemization, with 20 years of experience.

Manuel Requena (manuel.requena@cttc.es) is a senior researcher at CTTC and responsible for the EXTREME Testbed of the Services as Networks research unit.

Lorenza Giupponi (lorenza.giupponi@cttc.es) holds a Ph.D. from UPC (2007). She is a senior researcher at CTTC and a member of the CTTC Executive Committee.

CALL FOR PAPERS

IEEE COMMUNICATIONS MAGAZINE

FEATURE TOPIC: THE EVOLUTION OF TELECOM BUSINESS, ECONOMY, POLICIES AND REGULATIONS

BACKGROUND

From the late 1970's onward, the global telecommunication industry has gone through several waves of regulatory and technological changes. Today, the very definition of telecommunication has changed. First, the traffic is no longer confined to voice and limited data services, but has become a fundamental pillar of the digital transformation that is affecting all aspects of our contemporary life. Second, the boundary line between telecommunication and information technology is becoming blurred. Third, the parties that offer products and services for consumers, businesses and industrial users operate under different regulatory regimes, ranging from strict regulations to no regulations at all. Fourth, the merging of the physical and virtual worlds through immersive technologies, open new social, cultural and business dimensions. Finally, the post-covid environment has opened new avenues for virtual workplaces (E-work).

This overall transformation is affecting the human experience, whether individual or collectively, in unprecedented ways. In this environment, network operators, component/system suppliers, but also regulators, business analysts and social scientists as well as philosopher are looking at ways to understand the evolving world in terms of models, economic systems, and policies.

Traditionally, the focus of papers published in the IEEE Communications Magazine has been mostly on the technical aspects, but there is an increasing awareness of the interplay between technologies and designs on one hand and societal, economic, and regulatory factors on the other. Telecommunication specialists may have tended to relegate these latter factors to others. This has led them to operate in an environment to the definition of which they do not contribute, beside new technical features or products. It is high time for engineers and technologists to be involved in the non-technical aspects of their industry in the 21st century.

The purpose of this Feature Topic (FT) is to expand the engineering horizon in this regard. Thus, this FT solicits articles that explore the myriads of issues concerning future telecommunication business models, corporate strategies, innovations in financing and economics, and policy and regulatory matters. Potential topics include, but are not limited to:

- The telecommunication industry's evolution in terms of future technologies and social environment (e-work, virtual workplace, etc.).
- Near-future services and products from the operator, vendor, customer and regulator's perspectives.
- Strategies for the deployment of new technologies (6G, virtualization, etc.).
- Impact of immersive technologies: extended realities (XR), augmented reality (AR), mixed reality (MR) and virtual reality (VR).
- Impact of big data, artificial intelligence (AI) and distributed ledger technology on security, trust management, and privacy protection.
- Effect of new business models and corporate strategies on research and development in the telecommunication industry.
- Competition (intra-industry and with new entrants), vertical integration and regulatory impact.
- Telecom finances, e.g. return on capital expenses, debt, and funding.
- Telecommunications policy and the evolution of regulatory frameworks.
- Managing innovations in telecommunication products and services.
- Telecommunications and entrepreneurship.

SUBMISSION GUIDELINES

Manuscripts should conform to the standard format as indicated in the Information for Authors section of the *IEEE Communications Magazine's* Manuscript Submission Guidelines. Please, check these guidelines carefully before submitting since submissions not complying with them will be administratively rejected without review.

All manuscripts to be considered for publication must be submitted by the deadline through Manuscript Central. Select the "FT-2212/The Evolution of Telecom Business, Economy, Policies and Regulations" topic from the drop-down menu of Topic/Series titles. Please observe the dates specified here below noting that there will be no extension of submission deadline.

IMPORTANT DATES

Manuscript Submission Deadline: 30 October 2022
Decision Notification: 15 December 2022
Final Manuscript Due: 1 January 2023
Tentative Publication Date: March 20233

GUEST EDITORS

Hashem Sherif
AT&T (retired), USA
hashem_sherif@icloud.com

Eva Ibarrola
Escuela Superior de Ingeniería de Bilbao, Spain
eva.ibarrola@ehu.eus

Kai Jakobs
RWTH Aachen University, Germany
kai.jakobs@cs.rwth-aachen.de

Duncan Sparrell
sFractal Consulting, USA
duncan@sfractal.com

Distributed Trust and Reputation Management for Future Wireless Systems

Dev P. Singh, Kevin W. Sowerby, and Andrew C. M. Austin

The authors propose a distributed, three-layer, trust-based hardware sharing scheme between operators that overcomes the limitations of a single-operator-owned monolithic network. Their system scales to tens of thousands of operators without requiring explicit contracts between them, or between operators and user devices.

Abstract

Trust lies at the center of the paradigm shift required to realize the ultra-dense networks needed by future radio communication systems. We propose a distributed, three-layer, trust-based hardware sharing scheme between operators that overcomes the limitations of a single-operator-owned monolithic network. Our system scales to tens of thousands of operators without requiring explicit contracts between them, or between operators and user equipment. User equipment in turn is free to requisition the services of any available hardware. This is achieved by abstracting the communication process as a transaction, and casting it within a distributed ledger technology framework paired with an efficient, fault-tolerant, distributed consensus protocol. A trust model associates a behavioral measure with each hardware device that signals its reliability, as well as its payoff. The proposed system offers multiple advantages for users, operators, and regulators.

Introduction

Radio communication systems of the future will move away from fixed infrastructure providers and static contracts, and toward multi-tenanted systems featuring actively negotiated terms of service. This setting will feature thousands of hardware providers, including traditional and non-traditional cellular operators. Network hardware and user equipment (UE), will freely interact with each other to set up, execute, and resolve communication tasks without any static prior contracts. Consequently, the reliability of UE, network hardware, and service provisioning will need to be actively considered in the setup and execution of communication tasks. In a traditional cellular system, trust management occurs through periodically revised static contracts. An assumption of complete trust holds during a contractual period. Such an approach is not suitable for ultra-dense massively shared network hardware (SNH) systems due to the significant overheads introduced by explicit contracts between all UE and SNH, as well as between SNH belonging to different operators. Additionally, the reluctance of rival operators to freely share information or make their proprietary systems public may invalidate the perfect trust assumption. Finally, repeat interactions between network hardware and UE in an ultra-dense system may not occur frequently enough to generate statistically significant reliability measures.

Therefore, UE and SNH need to resolve terms of interaction within their local context, based on each other's uncertain or incomplete information. This is precisely the type of conditions under which trust enables distributed decision making within social systems. Trust within such settings is contexualized to a given activity, applied to direct interactions, and based on the discretion of the trusting party. Consequently, social-trust-based measures need to correlate the trust rating of a device with its actual performance at fulfilling its stated role. For instance, a UE in a shared setting is expected to pay for the services it consumes, whereas an SNH device is expected to adhere to the terms of the service level agreement (SLA). Experiences and outcomes perform a central role in trust-based decision making. Therefore, a trust-based system needs to record the history of relevant interactions. Furthermore, in order for these records to form a meaningful basis, they must meet information security requirements such as consistency, availability, and immutability.

A massive SNH system that replaces fixed subscriptions between users and operators with dynamic, locally initiated, trust-based contracts mandates a distributed notion of trust. This is to ensure that the computation, update, and propagation of trust occurs without centralized coordination or pre-trusted relationships/trusted third parties.

We propose a trust-based SNH system that meets the requirements of distributed setup, execution, and control of network activity. Our system consists of a distributed ledger technology (DLT) [1] component, paired with a Byzantine fault-tolerant (BFT) [2] consensus protocol, and a distributed trust model. The trust model uses a combination of behavioral and commodified trust to translate native DLT trust structures into communication system equivalents.

Our system offers several advantages for SNH operators and UE, including infrastructure independence, dynamic contracts, and a practical, low-cost pathway to network densification. Additionally, it provides regulators with a distributed framework to implement policies related to radio resource management, fair use, and energy efficiency.

Related Work

Existing work adopts one of two main approaches to addressing the high-frequency, ultra-dense network (UDN) deployment problem. The first of these is framed as the Network Embedding Problem (NEP) [3]. NEP invokes software defined control and service-oriented architecture to abstract hardware resources into groups of network functions. Each such logical partitioning, termed a network slice [3], is then allocated to service incoming user requests. The second widely used approach focuses on the perfor-

Digital Object Identifier:
10.1109/MCOM.002.2100959

The authors are with the University of Auckland, New Zealand.

0163-6804/22/$25.00 © 2022 IEEE

mance of infrastructure sharing in UDNs. Stochastic geometry is used to model the distribution of user equipment and operator hardware. Performance metrics such as signal-to-interference-plus-noise ratio (SINR), and outage probability derived from these models [4] guide deployment.

Both of these approaches assume single-tenant systems and focus on sharing of resource blocks rather than hardware. Consequently, most works exploring resource sharing are formulated in a centralized setting. Our work employs hardware sharing, extended to a multi-tenant setting, featuring thousands of operators committing network hardware, and without fixed contracts between UE and SNH. Additionally, the allocation, pricing, and management of shared network functions is implemented in a fully distributed manner. Recent works have looked at DLT-based resource sharing in mobile edge computing [6], and Industrial Internet of Things [5] settings, focusing on obfuscation of identity and routing topology respectively.

Distributed systems have been extensively researched. This has produced a rich family of system specifications and protocols. However, until the advent of DLT, most such systems were relegated to tightly constrained academic or private settings. DLT represents the only production-grade, large-scale, distributed, fault-tolerant system. The central idea originates from multi-party computation (MPC) [7] which uses state machine replication (SMR) to synchronize the states of multiple servers concurrently responding to service requests from clients. The service is abstracted as a state machine replicated across the servers. Paxos [8] style protocols were among the first to solve the SMR distributed consensus problem by invoking the atomic broadcast [2] primitive.

An SMR service is termed fault-tolerant if it progresses despite crashes or corruptions among a minority quorum of replicas. Such protocols proceed in rounds of message passing such that at the end of a round, all honest parties are guaranteed to commit to the same set of values. In the case of DLT systems, the value being committed to is a batch of transactions, and all correct participants commit by appending the batch to their local ledger as a new DLT block. Each round consists of protocol participants broadcasting their proposals, voting on each other's proposals, and using majority vote to decide. The voting mechanism is generally implemented using threshold cryptosystems [9, 10].

DLT systems generate leaderless agreement among nodes on the state of the DL without needing a trusted third party, such as a bank or central server [11]. Such systems are thus well suited to a massively shared, decentralized communication system featuring thousands of traditional and non-traditional operators. Bitcoin, the pioneering DLT system introduced proof-of-work (PoW)-style protocol [11] to solve consensus. PoW protocols have unacceptably high energy consumption due to the computational cost associated with securing the public or permissionless DL against malicious actors. However, in a permissioned DLT network, participants are not anonymous, allowing PoW to be substituted by more efficient, provably secure protocols from the field of fault-tolerant distributing computing [12, 13].

Distributed consensus protocols adopt different notions of reliability and are classified accordingly. A protocol is termed synchronous if message delivery occurs within bounded time. Asynchronous variants feature no external timing reference, and employ randomization protocols such as coin tossing [2] to measure progress. Messages in a radio communication system may be lost, delayed, or arrive out of order due to noise, interference, channel state variability, and other sources of nondeterminism. Therefore, asynchronous agreement protocols are better suited to specify such systems.

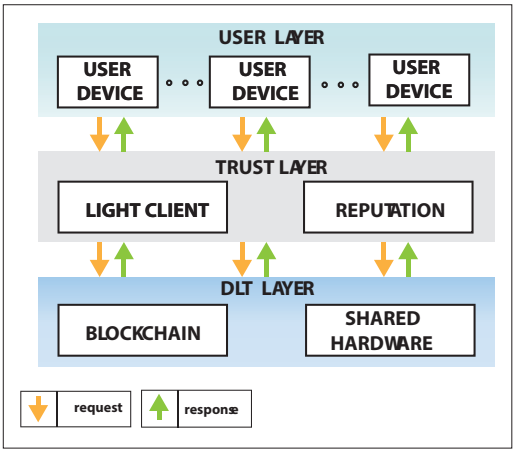


FIGURE 1. Three-layer model of the proposed system.

System Model

Our proposed system is abstracted as a three-layer model, as shown in Fig. 1, with the lower layers responding to service requests from upper layers. The lowest DLT layer comprises the blockchain framework. A UE in the system interacts with the shared hardware through the DLT light client and the reputation model. The DLT light client manages a UE's past transactions, available balances, transaction composition, and third-party interactions. The trust model is employed as a decision making device by both SNH and UEs.

Figures 2 and 3 illustrate the different stages of communication in our system, as an interaction between a UE and an SNH. A UE initiates the process whenever it needs to execute a communication task over the network. Compared to traditional dedicated hardware, UE-SNH interactions involve several additional stages concerned with setup, execution, and control. Trust management in our system involves representing, computing, updating, and storing trust. Trust is represented as a behavioral measure correlated to an SNH device's past record of fulfilling its stated function. This takes the form of a reputation rating assigned to each device. Every interaction between an SNH and a UE, encoded as a DLT transaction, triggers a change in SNH reputation. Updated reputations are included within the DLT transaction and are verified via the consensus protocol, which adds new transactions to the DLT ledger. The distributed nature of the ledger ensures that trust values reliably propagate through the system and are readily available. An SNH device in our system, acting as a DLT consensus node, is capable of self-managing all aspects of their operation. A software-based formulation of this node ensures that SNH devices can easily be switched between shared and private modes at little cost to the operator. This flexibility reduces the barrier to entry for an SNH device wanting to join the system, thereby enhancing scalability.

An SMR service is termed fault-tolerant if it progresses despite crashes or corruptions among a minority quorum of replicas. Such protocols proceed in rounds of message passing such that at the end of a round, all honest parties are guaranteed to commit to the same set of values.

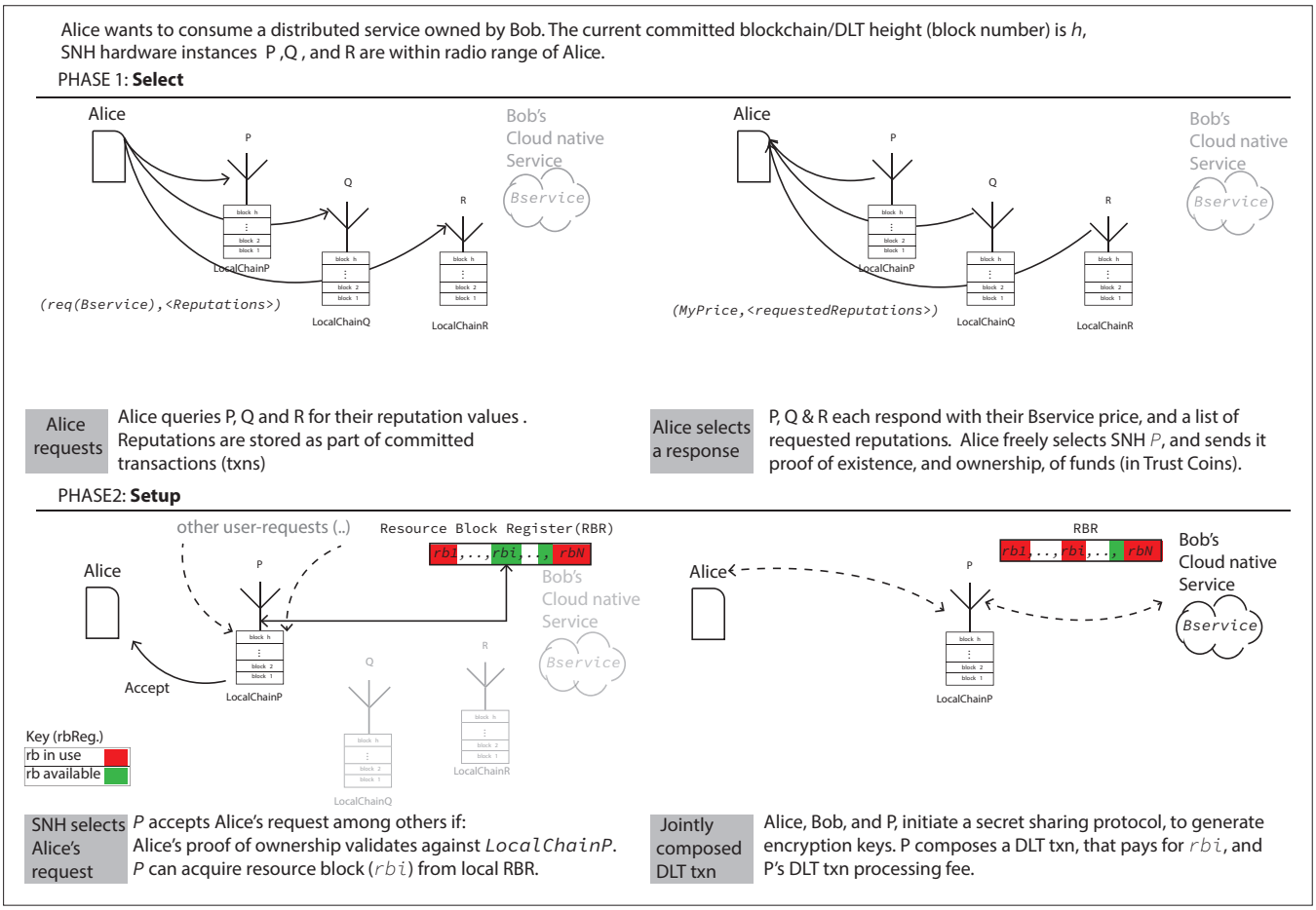


FIGURE 2. Phases involved in fulfilling a user request.

Trust Model

Trust modeling has traditionally occupied the realm of soft security as an alternative to hard security measures based on cryptography. Therefore, trust has often been applied to energy or bandwidth constrained systems such as wireless sensor networks. However, technological advances along with higher bandwidths accessible at millimeter-wave (mmWave) and higher spectrum bands enable applying both trust- and cryptography-based approaches to managing wireless communication systems. The key insight of our system lies in securing trust management processes through cryptography.

Our system absorbs trust interactions within DLT transactions. A transaction in the system encodes identities, obligations, and outcomes of these interactions. A trustor applies an application-specific trust function over some subset of transactions. The threshold for trust set by an agent is based on subjective factors, such as their experience with the system, and the importance they attach to the task at hand. Our system adapts the token of a traditional DLT called a TrustCoin (TC) to serve as the native cryptocurrency. Furthermore, by pricing network services based on the provider's reputation, a transaction in our system explicitly records reputation information.

Communication System

We assume that service discovery and provisioning occurs over a software defined networking (SDN) substrate [14] employing hardware virtualization techniques such as a cloud native network function (CNF) [3] to define and compose network services. CNF-based services can be viewed as a sequence of tethered functions that take as input the service request and produce the final desired output. Depending on granularity of control defined by the request, checkpoints can be established along multiple-input/output interfaces along this sequence. These in turn may be collaboratively monitored by transaction participants using threshold encryption schemes such as Threshold Public Key Encryption [2]. Such schemes are employed by the BFT consensus protocol used by the system; therefore, setting them up for this purpose does not add computational cost.

We assume the existence of a dynamic spectrum management policy. Following the principle of frequency reuse, we divide the network into disjoint clusters of SNH devices.Each such set is served by an exclusive resource block register (RBR), as shown in Fig. 3, thus reducing radio spectrum management to a single cluster and its allocated RBR. We also make the following modifications to the DLT transaction setup, composition, and validation. Transaction setup includes a check of the associated RBR for resource block availability. The transaction proceeds only if this condition is met. The owner of the relevant resource block (e.g., the cellular operator) then adds a new output to a proposed transaction that pays them the fee associated with the use of the resource. This fee is paid regardless of the outcome of the underlying service being transacted. Therefore, every proposed transaction is committed. Associating this cost with every proposed transaction prevents denial of service (DoS)-type attacks.

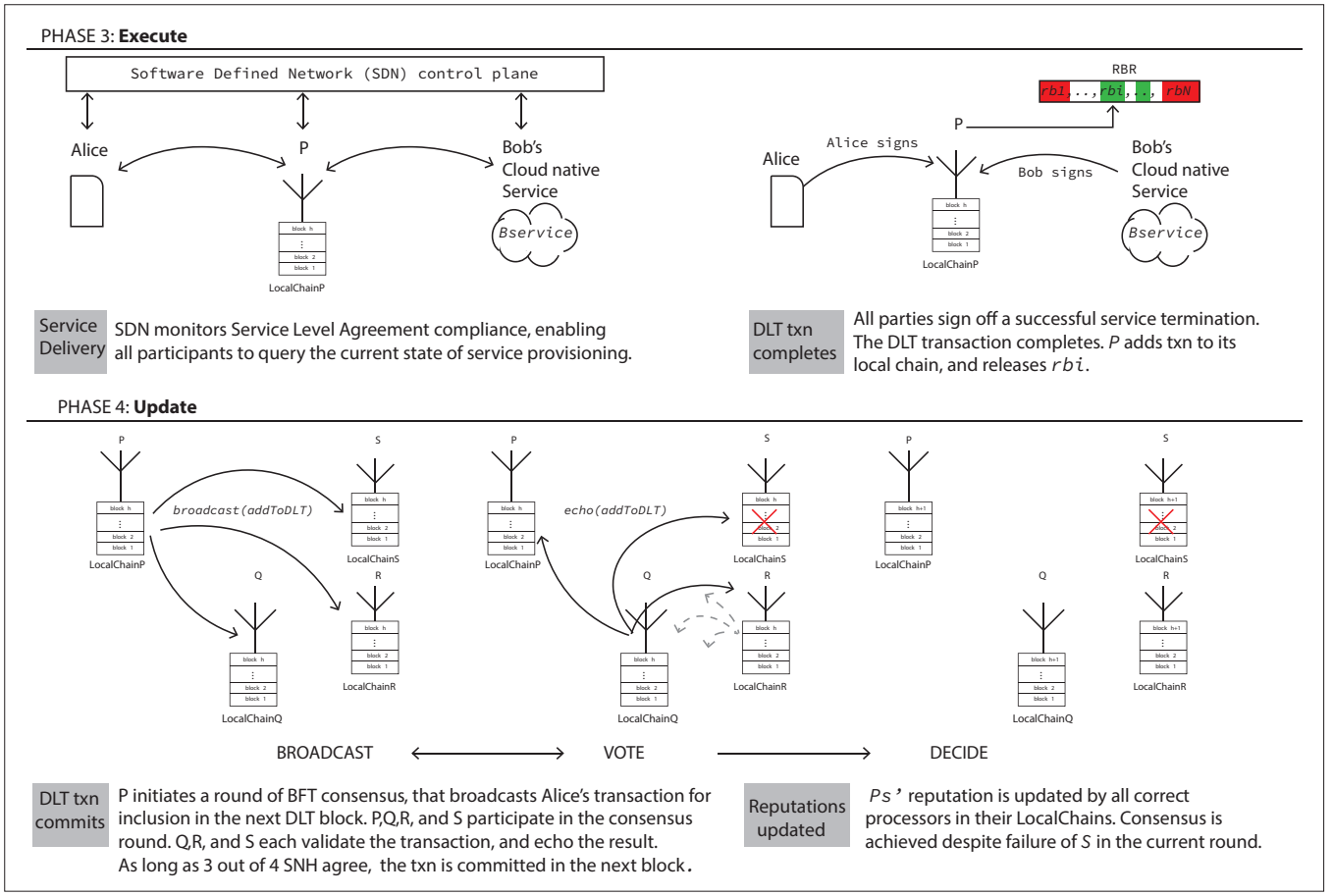


FIGURE 3. Phases involved in adding completed transactions to the ledger.

DLT Framework

DLT systems may be described using a four-layer architecture comprising the user interface layer, application programming framework, compute layer, and consensus layer. UEs submit requests generated over the user interface to the application programming framework, which applies a semantic interpretation. This semantic description is validated by the compute layer. Successfully validated transactions are forwarded to the consensus layer, which batch processes them for inclusion in the ledger.

Desirable features of DLT systems are derived from the mathematically provable security properties of cryptography protocols. Trust processes in our system are defined over such constructions, thereby extending their guarantees into the trust realm.

Each DLT node maintains an independent copy of a ledger of network activity. A consensus, or agreement protocol, is a distributed mechanism for nodes to synchronize the states of their individual ledgers. A BFT [12] consensus protocol progresses despite a quorum of protocol participants displaying malicious behavior. Such protocols comprise a propose phase followed by an accept phase. During the propose phase, participating nodes invoke the Reliable Broadcast primitive to disseminate their proposals, which are subsequently voted on for inclusion in their local ledgers. Broadcast and voting schemes are implemented over quorums of correct processors using threshold cryptography [2, 10].

Each SNH device is incentivized for participating in a consensus round by receiving a quantity of TC proportional with both their long-term reputation and their participation in the current round. This is in line with each SNH device being able to act purely as a service delivery node, consensus node, or both.

In PoW-style protocols, all the newly minted tokens are awarded to one miner; however, the payoff structure of our protocol needs to reflect its collaborative nature. In our system, SNH devices compute each other's share of the block reward based on the correctness of protocol messages generated by each participant, which also serves as a measure of reputation. Each participating SNH generates its own coinbase transaction [1]. Participating SNHs undertake an additional round of messaging to generate agreement over the independently computed coinbase transaction. Therefore, trust computation, storage, and update are folded into DLT processes and data structures.

Motivating Example

As described graphically in Fig. 2, Alice wants to avail of *Bservice* provisioned by Bob. Alice first sends a query to discover SNH devices within network range. SNH devices *P*, *Q*, *R*, and *S* respond with their credentials, which include their reputation and cost for delivering *Bservice*. The cost is stated in units of TC, while the reputation is a normalized numerical score. Alice reviews the responses and finds *P*'s terms suitable, and signals its acceptance. Alternatively, Alice may decide that none of the SNH devices suit her current need, and can abort or postpone the request. We assume the low-bandwidth control signaling occurs over publicly licensed spectrum bands.

This work presents a new paradigm for pervasive radio communication systems of the future. However, it requires a sea change in the way networks are built, operated, and viewed by traditional carriers.

SNH *P* reviews Alice's user request among the others it has received, and initiates a secret sharing protocol [9] with Alice and *Bservice* to generate transaction-specific keys. These secret shares are combined to generate a digital signature to prove TC ownership, and collaboratively monitor progress. Software defined controls monitor the status of *Bservice* and communicate it to transacting parties. All parties sign the successful completion of *Bservice*, and *P* receives its payment. *P* stores the completed transaction in its local buffer, scheduled for inclusion in the DL.

During a given round, each participating SNH runs one instance of the agreement protocol for every proposal. Participant *P* deems a round complete when two-thirds or more of these instances terminate.

System Validation

We validate the system by generating and model checking a formal specification. The specification is a mathematically precise description of system behavior. A model checker generates all possible system traces resulting from the specification to determine whether it violates any specified property.

We specify our system using Leslie Lamport's Temporal Logic of Actions (TLA) [15]. TLA has been developed for distributed and concurrent systems, and is able to pick up subtle bugs missed by traditional verification tools such as Monte Carlo simulations and unit testing. A state-machine abstraction of the system is described using set theory and first-order logic. A state is a unique assignment of values to system variables. An action enables transitioning to a new state by acting on state variables. The property that needs checking is described as a Temporal Logic formula, and forms an invariant of the system: a formula that must be satisfied at every state along all sequences of states. Our specification consists of three sets of functionally distinct processes, representing UEs that propose values, SNHs that validate these proposals, and those that commit them to the DLT ledger.

We model consensus with a Byzantine version of the popular Paxos [8] protocol. The protocol is extended to account for tokenization and protection against double spending [1]. Our specification satisfies safety, and correctness properties [15], such as not allowing double spending of TC and ensuring that the local unspent transaction output [1] of all correct processors are consistently edited.

Open Issues and Challenges

Our proposed system needs to coordinate its actions across disparate components of a radio communication system. The deployed dynamics of radio resource management and software defined controls may necessitate tuning specified system parameters, such as setting block and transaction sizes based on network throughput and latency. Reconfiguration of the consensus protocol to manage changes to membership, without disrupting availability of the system, remains an open challenge.

Even though model checking our system specification validates its correctness, it is necessarily carried out over a finite model size. Therefore the impact of potentially limitless scaling remains unclear, especially given the well-known inability of BFT consensus protocols to scale in the presence of faults. This is mitigated by leveraging the frequency-dependent partitioning of radio systems to create clusters of consensus groups.

Conclusion

This article introduces a trust-based framework to implement a massively shared, multi-tenant wireless communication system. Our system overcomes the challenges of deploying ultra-dense networks by dramatically expanding the class of operator admitted into the system. We replace rigid trust with flexible terms negotiated independently between transacting UE and SNH. Trust in our system is derived from the actual record of network activity enshrined as DLT transactions. Therefore, trust functions may be freely defined over arbitrary subsets of transactions. We have outlined the operational, economic, and technical components of the framework and validated its safety properties. This work presents a new paradigm for pervasive radio communication systems of the future. However, it requires a sea change in the way networks are built, operated, and viewed by traditional carriers.

References

[1] A. M. Antonopoulos, *Mastering Bitcoin: Programming the Open Blockchain*, O'Reilly Media, 2017.

[2] A. Miller *et al.*, "The Honey Badger of BFT Protocols," *Proc. 2016 ACM SIGSAC Conf. Computer and Commun. Security*, Vienna, Austria, Oct. 2016, pp. 31–42.

[3] S. D. A. Shah, M. A. Gregory, and S. Li, "Cloud-Native Network Slicing Using Software Defined Networking Based Multi-Access Edge Computing: A Survey," *IEEE Access*, vol. 9, 2021, pp. 10,903–24

[4] R. Jurdi *et al.*, "Modeling Infrastructure Sharing in mmWave networks with Shared Spectrum Licenses," *IEEE Trans. Cognitive Commun. and Net.*, vol. 4, no. 2, 2018, pp. 328–43.

[5] H. Yang *et al.*, "Blockchain-Enabled Tripartite Anonymous Identification Trusted Service Provisioning in Industrial IoT," *IEEE IoT J.*, 2021.

[6] H. Yang *et al.*, "Distributed Blockchain-Based Trusted Multidomain Collaboration for Mobile Edge Computing in 5G and Beyond," *IEEE Trans. Industrial Informatics*, vol. 16, no. 11, 2020, pp. 7094–104.

[7] M. Pease, R. Shostak, and L. Lamport, "Reaching Agreement in the Presence of Faults," *J. ACM*, vol. 27, no. 2, 1980, pp. 228–34.

[8] L. Lamport, "The Part-Time Parliament," *ACM Trans. Comp, Sys,*, vol. 16, no. 2, 1998, pp. 133–69.

[9] D. Boneh, M. Drijvers, and G. Neven, "Compact Multisignatures for Smaller Blockchains," *Int'l. Conf. Theory and Application of Cryptology and Info. Security*, Brisbane, Australia, Dec. 2018, pp. 435–64.

[10] C. Cachin *et al.*, "Asynchronous Verifiable Secret Sharing and Proactive Cryptosystems," *Proc. 9th ACM Conf. Comp. and Commun. Security*, Washington, DC, Nov. 2002, pp. 88–97.

[11] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," Cryptography Mailing List, Oct. 2008; https://bitcoin.org/bitcoin.pdf.

[12] G. Bracha, "Asynchronous Byzantine Agreement Protocols," *Info, and Computation*, vol. 72, Jan. 1987, pp. 130–43.

[13] M. Ben-Or, B. Kelmer, and T. Rabin, "Asynchronous Secure Computations with Optimal Resilience," *Proc. 13th Annual ACM Symp. Principles of Distributed Comp.,* New York, NY, Aug. 1994, pp. 183–92.

[14] ITUR-WP5D, "Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s)," ITU Tech. Rep., 2017.

[15] K. Chaudhuri *et al.*, "Verifying Safety Properties with the tla+ Proof System," *Proc. 5th Int'l. Conf. Automated Reasoning*, Springer-Verlag, 2010, p. 142–48.

Biographies

Dev P. Singh [M'19] is currently pursuing a doctorate with the Department of Electrical, Computer, and Software Engineering (ECSE) at the University of Auckland.

Kevin W. Sowerby [SM'03] Professor Sowerby currently serves as Head of ECSE at the University of Auckland.

Andrew C. M. Austin [M'10] is currently a senior lecturer in the ECSE Department at the University of Auckland.

CALL FOR PAPERS

IEEE COMMUNICATIONS MAGAZINE

FEATURE TOPIC: FUTURE TRENDS IN FOG/EDGE COMPUTING AND NETWORKING

BACKGROUND

Over the past decade, cloud computing has played a dominant role in supporting the applications we rely on today. Mobile networks have been acting mostly as communication pipes connecting users to the cloud and with each other. As we evolve toward the Internet of Things (IoT), our 5G/6G and future mobile networks must support a much wider range of applications, including vehicular networking, automated manufacturing, smart cities, drones, smart grids, e-health, and the many emerging AI-enabled applications such as Virtual Reality (VR) and Augmented Reality (AR). The cloud computing plus communication pipe model is no longer adequate for supporting these emerging applications. For example, many IoT applications cannot tolerate the delays incurred by cloud computing. The endpoints are creating a vast and ever-growing amount of data that needs to be processed locally because sending all the data to the cloud will often be infeasible due to network bandwidth constraints and regulatory restrictions. Connecting every device directly to the cloud can often be impractical due to limited resources on the devices, software and management complexity, limited network agility and cognition, and system scalability. In such scenarios, users will desire local services. Many resource-constrained devices will also require local services to help perform many tasks that they cannot perform by themselves. Such tasks may range from computational-intensive user applications to security tasks that require heavy processing or information that the resource-constrained devices do not have. Future mobile networks will also require computing capabilities inside or close to the radio access networks (RANs) to enable advanced networking capabilities, such as establishing radio connections more timely and adjusting radio channel coding dynamically in response to changing user needs and communication environments, and allow user applications to be hosted in the RANs that are closer to the users.

These and the many other new requirements call for a new computing paradigm – fog/edge computing and networking. Fog/edge computing technologies envision an open horizontal architecture for distributing functions (from computing to storage to control and to networking functions) closer to users, not just to any specific type of network edge device but anywhere along the cloud-to-thing continuum that can best meet user requirements. Fog will integrate with the cloud to provide a seamless end-to-end computing platform along the cloud-to-thing continuum. Fog/edge services and user applications can be deployed anywhere along this computing continuum. The same function or application can be deployed and subsequently moved anywhere along the continuum, in the cloud, the fog, and even onto the endpoints, to best meet user requirements. Computing resources distributed along the cloud-to-thing continuum can be pooled together to support a user application. Fog/edge nodes will work autonomously when connectivity to the cloud is unavailable. They can also collaborate with each other to carry out tasks for the users.

Fog/edge technologies will play key roles in future computing and networking systems. Researchers investigate fog/edge computing and networking technologies to optimize of resources that are virtualized, pooled, and shared unpredictably. Fog networking revisits the role of clients in network architectures, more than just an end-user device, but also as an integral part of the control plane that monitors, measures, and manages the network. This is rewriting the traditional practice of using heavy-duty and dedicated network elements for network measurement and management fog/edge computing & networking combine the study of mobile communications, distributed systems, and big data analytics into an exciting new area.

With fog/edge computing and networking technologies, many new emerging services (such as V2V in Vehicular Telematics Services, Autonomous Car, Industry 4.0 and e- Healthcare Services) could be realized and implemented easily and economically. It could be also served as core engine to enable many Services in Internet of Things (IoT) applications. Vertical markets and applications will be critical for 5G/6G systems. There are opportunities in applying fog/edge technologies to facilitate the operations of vertical applications with integrated computing and communications design.

More recently, edge AI has emerged as the next frontier and a cornerstone for future intelligent networks. Edge AI has multiple levels. At the base level, edge devices use AI/ML models created somewhere else (typically in the cloud) to tackle complex tasks. At the top level, edge devices learn from their local data ("edge learning") to help create (train) AI/ML-based network functions and user applications. Edge learning is becoming necessary and essential because sending the massive amount of data created at the network edge to the cloud for ML model training is increasingly impractical.

This Feature Topic (FT) is aimed to cover a wide variety of recent advancement and future directions on fog/edge computing, including trends towards 6G fog/edge, cutting-edge fog/edge research contributions, experiments and performance of fog/edge computing systems, challenges, and opportunities for fog/edge, novel business models, and killer applications. We welcome viewpoints and contributions from academia and industry. The topics of interest for this Feature Topic include, but not limited to, the following:

- Visions toward future fog/edge evolution
- Fog/edge technologies for 6G
- Fog/edge based IoT services
- Fog/edge computing and networking for mission-critical services
- Edge AI and edge learning
- Management and orchestration for fog/edge systems
- Data analytics and machine learning in the fog/edge computing environment
- Security and trust in fog/edge systems
- Key organizations or consortia of fog/edge activities.
- Standards and future direction of fog/edge systems
- Experiences sharing of fog/edge testbeds and deployment.
- Fog/edge platform for vertical industries (e.g. manufacturing, transportation)

SUBMISSION GUIDELINES

Manuscripts should conform to the standard format as indicated in the Information for Authors section of the *IEEE Communications Magazine's* Manuscript Submission Guidelines. Please, check these guidelines carefully before submitting since submissions not complying with them will be administratively rejected without review.

All manuscripts to be considered for publication must be submitted by the deadline through Manuscript Central. Select the "FT–2221/Future Trends in Fog/Edge Computing and Networkings" topic from the drop-down menu of Topic/Series titles. Please observe the dates specified here.

IMPORTANT DATES

Manuscript Submission Deadline: 15 December 2022
Decision Notification: 15 April 2023
Final Manuscript Due: 1 May 2023
Tentative Publication Date: Mid-2023

GUEST EDITORS

Hung-Yu Wei (Lead Guest Editor)
National Taiwan University, Taiwan
hywei@ntu.edu.tw

Tao Zhang
National Institute of Standards and Technology, USA
tao.zhang@nist.gov

Russell Hsing
National Chiao Tung University, Taiwan
thsing@cs.nctu.edu.tw

Doug Zuckerman
Peraton Labs, USA
d.zuckerman@ieee.org

Future Directions for Wi-Fi 8 and Beyond

Ehud Reshef and Carlos Cordeiro

The authors provide an overview of Wi-Fi generations until Wi-Fi 7, and then discuss future market requirements, drivers, and technology challenges in meeting them.

Abstract

Generation after generation, Wi-Fi has experienced tremendous growth in both market penetration and technical capabilities. Wi-Fi 6/6E is the latest generation of the technology with commercially available products. In the meantime, the industry is already developing Wi-Fi 7, based on features defined in IEEE 802.11be. As the feature set of Wi-Fi 7 becomes well defined and product development is underway, it is essential to take a step back to analyze what the usages and technology drivers of the next 5–10 years will be. With this understanding and the expectation that Wi-Fi 8 could become a market reality in the 2027/2028 timeframe, we can anticipate technology candidates that can be part of Wi-Fi 8. In this article, we provide an overview of Wi-Fi generations until Wi-Fi 7, and then discuss future market requirements, drivers, and technology challenges in meeting them. We then introduce several possible technical directions for Wi-Fi 8 along different performance vectors, including throughput, capacity, and device density, as well as improved support for quality of service, latency and reliability, and operation in new frequency bands.

Introduction

Wi-Fi technology has been delivering high-performance wireless networking capabilities for more than 20 years. Wi-Fi enabled major market milestones involve the transition from stationary computing to the current hyperconnected world, such as the launch of Intel's Centrino program in 2003 and the first Apple iPhone in 2007.

Operating in unlicensed frequency bands and designed with an open ecosystem in mind, Wi-Fi technology has delivered significant innovations generation after generation, consistently pushing the performance envelope while enabling the design and introduction of multiple cost-effective user devices and access points into the market. Peak Wi-Fi data rates have exploded from the mere 1 Mb/s in the original 1997 IEEE 802.11 standard to almost 10 Gb/s (10,000×) with currently shipping Wi-Fi 6 technology based on IEEE 802.11ax, with laptops supporting data rates in excess of 2 Gb/s. This, in turn, allowed users to benefit from gigabit to the home broadband access services, and to rely on robust and secure cloud computing for business and personal computing. Given the high data rates and relative low cost of integration into devices, the technology has become ubiquitous and is the main vehicle of connecting to the Internet for many people. Indeed, over 63 percent of mobile data is offloaded to Wi-Fi, creating a $3.3 trillion (2021) [1] economic value.

To address emerging usages, such as wireless augmented/virtual reality (AR/VR), the Industrial Internet of Things (IIoT), and support the more demanding requirements originating from network evolution toward edge computing, artificial intelligence (AI), and virtualization, the industry is currently developing the next generation of Wi-Fi technology, under the IEEE 802.11be task group. 802.11be, expected to be the core of Wi-Fi 7 [2], brings with it a new paradigm shift for Wi-Fi, moving from a single connection over a single channel toward a collaborative multi-link operation that will allow applications to obtain greater capacity and lower latencies.

When looking beyond the horizon, it is easy to see that the days of extending the performance of Wi-Fi technology by relatively simple means such as doubling the channel bandwidth or extending the operation through use of spatial reuse technologies are over. Therefore, it is expected that future Wi-Fi generations will need to focus on other means or exploit new frequency bands to enhance performance. One clear need is to further reduce data delivery latency and jitter to optimize use cases such as AR, VR, and Industry 4.0. It is therefore expected that with future Wi-Fi generations, starting with Wi-Fi 8, means to improve spectral efficiency and reduced latency will become major focus areas for development.

This article starts with a review of the generational evolution that the Wi-Fi technology has gone through since Wi-Fi 4 (802.11n), with special focus on Wi-Fi 6, 6E, and Wi-Fi 7. Following that, the article reviews future Wi-Fi usages and the related technical challenges the industry is facing. Finally, it introduces some possible technical directions for future Wi-Fi standards such as Wi-Fi 8 and beyond.

Wi-Fi Generations

Wireless local area network (WLAN) technology was originally standardized in IEEE 802.11-1997. Branded as Wi-Fi, the technology we know today has been extensively modified through the 802.11a-1999, 802.11b-1999, and 802.11g-2003 amendments to support direct sequence spread spectrum operation in the 2.4 GHz industrial, scientific, and medical (ISM) band, and orthogonal frequency-division multiplexing (OFDM) operation in either of the 2.4 GHz or 5 GHz unlicensed bands. Intel unleashed the power of Wi-Fi in 2003 with the launch of its first mobile PC platform under the Centrino brand [3], a time when wireless Internet connectivity was beginning to be deemed a critical feature and a prerequisite for "unwiring the PC."

When the Wi-Fi Alliance created the Wi-Fi generation branding in 2019 alongside the intro-

Digital Object Identifier:
10.1109/MCOM.003.2200037

Ehud Reshef is with Intel Corporation, Israel; Carlos Cordeiro is with Intel Corporation, USA.

0163-6804/22/$25.00 © 2022 IEEE

duction of Wi-Fi 6 technology [4], it decided to forego branding 802.11a, b, and g each with a specific Wi-Fi generational number and decided to start the concise generational branding from 802.11n, the current baseline, branding it as Wi-Fi 4. Figure 1 shows the evolution of Wi-Fi technology to Wi-Fi 6. Wi-Fi 4 was the first Wi-Fi technology to support both the 2.4 GHz and 5 GHz bands, used an updated OFDM modulation scheme, and introduced three key innovations: the use of wider channel bandwidth with 40 MHz bonded channels, an improved forward error correction scheme using low density parity check (LDPC) codes, and the introduction of multiple-input multiple-output (MIMO) communications.

The next generation of Wi-Fi, recently rebranded as Wi-Fi 5, was based on IEEE 802.11ac and introduced to the market in 2013 [5]. Intentionally only defined to operate in the 5 GHz band to promote Wi-Fi operation in this much larger frequency band, Wi-Fi 5 more than doubled the achievable data rates through use of wider channel sizes (e.g., 80 MHz) and the introduction of a 256-QAM (quadrature amplitude modulation). Wi-Fi 5 also increased the maximum number of usable spatial streams to 8, although only 4 spatial stream operation was eventually certified, and enabled the use of beamforming by aligning on a single scheme of explicit feedback beamforming. In 2016, Wi-Fi 5 was enhanced with the introduction of 160-MHz wide channels and downlink multi-user MIMO (MU-MIMO). This enabled the introduction to the market of gigabit Wi-Fi products (e.g., Intel Wireless-AC 9260) that supported 1.7 Gb/s PHY data rates and could deliver actual user data rates greater than 1 Gb/s.

Quality of service (QoS) delivery has also improved with every generation of Wi-Fi, starting with the requirement that every device support prioritized channel access with IEEE 802.11e. Moreover, capacity improvements with the addition of larger numbers of spatial streams, wider channels, and frame aggregation have also helped substantially improve Wi-Fi QoS. For example, latencies have improved by at least one order of magnitude.

Further to extending peak performance with the transition to Wi-Fi 4 and Wi-Fi 5 technology, there is still need for lower cost and power Wi-Fi devices, mostly in the IoT segment. IoT devices can be delivered with optimized cost by implementing a subset of the standardized capabilities, for example, by only supporting 20 MHz wide channels or by limiting Tx power to below the regulatory allowed maximum. In many cases, it has been up to the vendors to decide what would be the specific subsets to be implemented in order to optimize their solutions to these markets.

Wi-Fi 6 and Wi-Fi 6E

In 2019, the next generation of Wi-Fi was introduced to the market as Wi-Fi 6. Wi-Fi 6 is based on the IEEE 802.11ax standard amendment and introduced a major paradigm shift for Wi-Fi technology.

Whereas all prior Wi-Fi generations relied on each device independently contending for channel access, Wi-Fi 6 introduced a trigger-based operation mode whereby an access point (AP) can trigger multiple stations (STAs) to transmit data concurrently. Trigger-based access was required to enable the introduction of another major capability

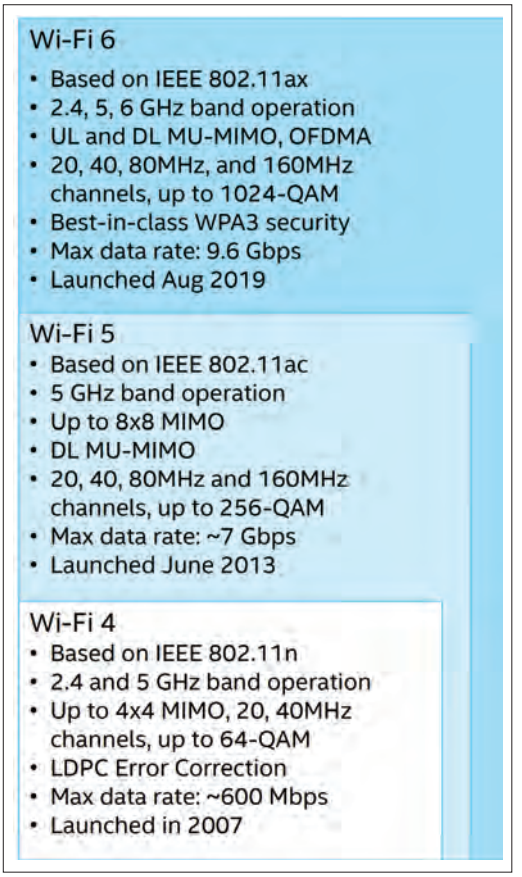


FIGURE 1. Wi-Fi generations.

— uplink multi-user operation. Other new features introduced with Wi-Fi 6 in 2019 were:

1. A new symbol format that increased symbol time and reduced OFDM carrier spacing by 4x, resulting with higher spectral efficiency, and increased multipath resiliency
2. Downlink (DL) and uplink (UL) orthogonal frequency-division multiple access (OFDMA), which enables multi-user operation on both DL and UL using OFDMA technology
3. Individual target wake time (TWT), which enables a STA to negotiate availability times with its AP to improve power consumption

The new access scheme, in conjunction with OFDMA support and other Wi-Fi 6 features such as spatial reuse and the underlying capacity enhancement features continued to reduce latencies and improve overall QoS delivered by Wi-Fi, especially in spectrum congested environments. Further to the technical specification, Wi-Fi 6 technology and certification added a dedicated certification profile for devices supporting only 20 MHz channels. This, in conjunction with certifying features such as Max BSS Idle period, have created a device profile better fitting the IoT market segment, where the key consideration is device cost and long battery life rather than peak performance.

Spectrum is the lifeblood of wireless technologies. One of the limiting factors for Wi-Fi to deliver gigabit data rates in dense environments is the limited amount of available unlicensed spectrum. In most countries around the world, the 2.4 GHz ISM band is only 78 MHz to 84 MHz wide, allowing use of up to three or four nonoverlapping 20 MHz channels. The band is further heavily used

The new access scheme, in conjunction with OFDMA support and other Wi-Fi 6 features such as spatial reuse and the underlying capacity enhancement features continued to reduce latencies and improve overall QoS delivered by Wi-Fi, especially in spectrum congested environments.

With MLO, Wi-Fi technology allows devices to benefit from coordinated operation on more than a single channel, achieve enhanced throughput (aggregating data at the MAC layer), reduced latency (transmitting data over any currently available pre-configured channel), and enhanced quality of service (QoS) (QoS-aware allocation of traffic to channels/links).

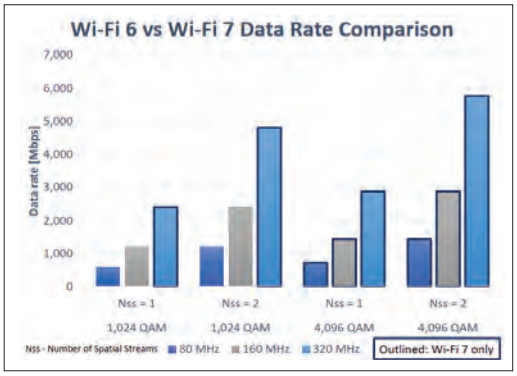


FIGURE 2. Wi-Fi 6 vs. Wi-Fi 7 data rate comparison.

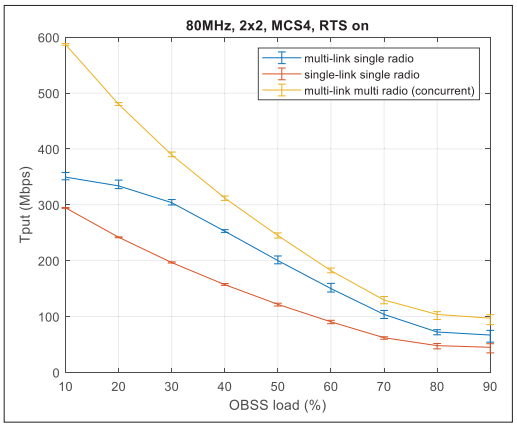


FIGURE 3. eMLSR, MLSR, and STR MLMR throughputs [10].

by Bluetooth and other unlicensed wireless devices. Moreover, the 5 GHz band only has about 480 MHz of unlicensed spectrum, and even that is limited in most countries to indoor operation or constrained by the requirement to avoid interference with weather and other radar signals.

As a result, the Wi-Fi industry has been advocating regulators to allocate new spectrum for unlicensed use. In a momentous decision, the FCC proposed allocation of 1.2 GHz of additional spectrum in the 6 GHz band (5925–7125 MHz) to unlicensed operation more than tripling the amount of spectrum available for Wi-Fi in 2018, with the final allocation ratified in 2020 [6]. Given the (justified) expectation that other countries would follow the FCC, the Wi-Fi ecosystem extended the IEEE 802.11ax specification to enable Wi-Fi operation in this new band. The current regulatory landscape is still evolving, with several countries following the United States and allowing unlicensed operation in the complete 6 GHz band (e.g., Korea, Brazil, Chile), while EU and several other countries have only allowed unlicensed operation in the lower 500 MHz of the band [7]. It should be noted that there are several countries, such as China, where 6 GHz unlicensed access is not allowed, at least for now. The extension of Wi-Fi technology into the 6 GHz band was introduced to the market in early 2021 as Wi-Fi 6E. Alongside defining the new 6 GHz channelization and operation, Wi-Fi 6E introduced several new features to optimize for the unique regulatory requirements of the band, and to optimize the operation in the band given no legacy Wi-Fi devices are allowed in it.

Finally, several 802.11ax features that were left out of the initial Wi-Fi 6 and Wi-Fi 6E certifications, such as uplink (UL) MU-MIMO, were introduced to the market in 2021 as a follow-on revision to the Wi-Fi 6 program [8]. Other features introduced with this release include power save enhancing features such as the use of TWT information elements that can be used to optimize individual TWT, and broadcast TWT that can simplify AP schedulers and use of TWT. It should be noted that TWT (individual or broadcast) could also be utilized to manage QoS across the Wi-Fi network.

Wi-Fi 7

With Wi-Fi 6 secured, the IEEE and Wi-Fi industry started work on the next generational upgrade under the IEEE 802.11be program, with expectations for products supporting Wi-Fi 7 technology based on 802.11be to be available in the market as early as 2024 [9].

With 802.11be already at draft D2.0 mid-2022, the key features of the technology are already known and include:

- Multi-link operation (MLO)
- Support of 320-MHz-wide channels
- 4096-QAM modulation scheme
- Operation in punctured channels (through the multi-RU operational mode)
- Triggered TxOp sharing procedure
- UL and downlink (DL) single-user and multi-user OFDMA and MIMO with up to eight spatial streams

Further to these features, Wi-Fi 7 is aiming to enhance QoS and deliver deterministic low latency through features such as rTWT and smart use of MLO, as discussed below.

As shown in Fig. 2, these features will enable a single 320 MHz link in the 6 GHz band to deliver up to 5.76 Gb/s over a two-spatial-stream link, and by aggregating a 160 MHz link in the 5 GHz band with a 320 MHz link in the 6 GHz band to deliver up to 8.64 Gb/s.

The key feature Wi-Fi 7 introduces is MLO. With MLO, Wi-Fi technology allows devices to benefit from coordinated operation on more than a single channel, achieve enhanced throughput (aggregating data at the medium access control, MAC, layer), reduced latency (transmitting data over any currently available pre-configured channel), and enhanced QoS (QoS-aware allocation of traffic to channels/links).

Wi-Fi 7 technology defines five different types of MLO devices [14]:

1. Multi-link single radio (MLSR) devices: These devices have a single radio interface and can either transmit or receive over a single link at a time. While these devices do not actually concurrently operate on multiple links, the underlying association is for all links, and there could be different schemes to enable MLO with these devices (e.g., using mutually exclusive per-link TWT windows).
2. Enhanced multi-link single-radio (eMLSR) devices: These devices can listen to more than one link at any given time, and through that always be available to either transmit or receive on more than one pre-configured (associated) link/channel.
3. Non-simultaneous transmit-receive multi-link multi-radio (NSTR-MLMR) devices: These devices can either transmit or receive (but not transmit and receive) concurrently on

more than one link.

4. Simultaneous transmit-receive MLMR (STR-MLMR) devices: These devices can transmit and/or receive concurrently on more than one link.
5. Enhanced MLMR (eMLMR) devices: These MLMR devices can dynamically reconfigure their spatial multiplexing capabilities by supporting multiple configurations of RF chains to links across the overall available set of RF chains. This could be beneficial in the case of an RF chain being able to support multiple different bands (e.g., a single RF chain can support either the 5 or 6 GHz band).

The underlying expectation is that APs will support operating in all flavors, whereas client devices (STAs) will support a subset of the operational modes. The specific value of eMLSR is that with modest additions to device cost and complexity, such devices can achieve much of the MLO value of full STR-MLMR devices. Figure 3 compares the performance of leading MLO flavors, represented as the achievable throughput for different channel loads,simulated as traffic on a different network on the same channel — hence as overlapping basic serving set (OBSS) load.

To achieve enhanced quality of service and deterministic low latency, 802.11be is introducing two key features:

- Traffic identifier (TID) to link mapping: With this feature, an AP can direct latency-sensitive traffic to a specific band/channel, while relegating "background" and other latency-tolerant traffic to a different channel/band. This feature would work best with STR-MLMR stations, since if used with single-radio stations, the restriction to work on a specific link (or subset of links) may create situations where link-constrained traffic would be delayed compared to when using the "all TIDs to all links" default mapping.
- Restricted TWT (rTWT): This feature defines TWTs where only targeted stations are expected to communicate with the AP. The standard aims to make these windows exclusive to targeted stations and thus ensure that their traffic meets their latency targets. The emerging specification, however, is somewhat hindered in this perspective by allowing stations that do not support the feature to ignore this exclusion.

In addition to features directly aiming to enhance QoS, Wi-Fi 7 MLO is expected to significantly reduce latency due to the increased data rates, and the expected reduced channel access time with eMLSR or STR dual radio.

It is further expected that Wi-Fi 7 devices may incorporate other QoS improvement capabilities that have been developed in IEEE and Wi-Fi Alliance since the launch of Wi-Fi 6 technology. One example is Wi-Fi QoS management [15], which defines features enabling DL and UL packet tagging and channel access treatment, as well as mapping of the DSCP field to 802.11 user priorities.

Future Usages and Technical Challenges

Looking ahead to the future, applications such as AR, extended reality (XR), and VR, cloud productivity and gaming, as well as emerging metaverse usages [11] will benefit from higher data rates and lower-latency network connections.

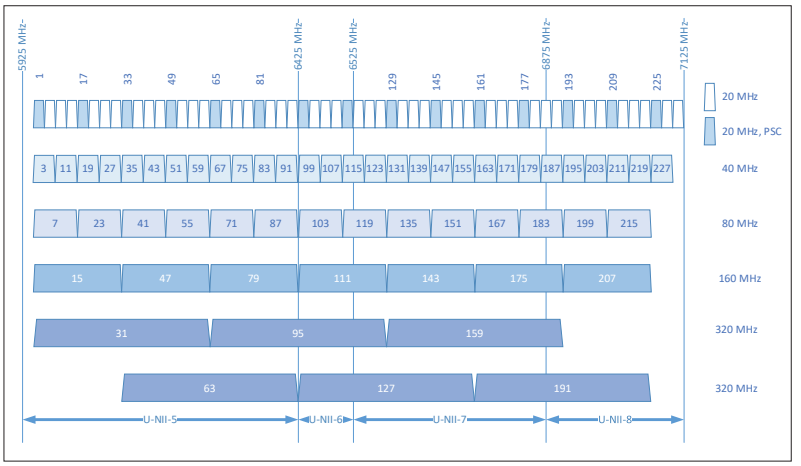


FIGURE 4. 6 GHz unlicensed band.

As Internet infrastructure is improving and cloud service providers (CSPs) deploy cloud edge servers in more locations, the end-to-end (E2E) latency contribution of Wi-Fi, while significantly improved in Wi-Fi 6 [12] and expected to be further improved in Wi-Fi 7, is becoming a dominant part of the E2E latency. This is especially true given Wi-Fi's promise of backward compatibility, which always allows for the situation where a single legacy connection can impact the overall network performance.

Another emerging wireless use case is the wireless factory, where there is need to deliver ultra-low latency coupled with high reliability, often without need for the very high data rates that Wi-Fi can deliver. While Wi-Fi can cater to such use cases, especially with rTWT, the current standard is not fully optimized for this scenario, preferring operation modes that account for legacy devices. With the growing importance of such deployments, there could be value in designing dedicated operational modes that rely on and optimize for "greenfield" and managed environments.

Another challenge for improving Wi-Fi performance lies in continuing to improve the achievable data rates. Improvements to date have been due to increased channel bandwidth, use of higher order modulation schemes, and use of more spatial streams. When looking ahead, we can see that there are challenges to continue to deliver real-world improvements across these three fronts:

- Channel bandwidth: Wi-Fi 7 is already specifying the use of 320 MHz channels. However, there are only six such channels available in the 6–7 GHz band across two overlapping grids, allowing for just three non-overlapping 320 MHz channels to be used (Fig. 4)
- Higher order modulation schemes: Wi-Fi 7 is specifying the use of a 4096-QAM (a.k.a. 4K-QAM) scheme. The next logical modulation scheme would be a 16K-QAM scheme. Regardless of the need to improve transmitter quality, the data rate increase due to such a transition would be a mere 16.66 percent (log2(16384)/log2(4096)), which is a very minor improvement, especially when considering the significant related implementation complexities.

Regardless of the eventual directions taken in existing Wi-Fi bands of operation, several WNG contributions mentioned the potential for Wi-Fi expansion into mmWave bands, and specifically into the unlicensed 60 GHz band. This band has up to 14 GHz of unlicensed spectrum in the 57–71 GHz range, more than 10× the available unlicensed spectrum below 10 GHz.

- More spatial streams. There are several different angles here:
  - –AP perspective: Wi-Fi 7 already defines operation of up to 16 spatial streams, although it is hard to envision a device with more than 8 antennas any time soon due to aspects such as form factor, cost, and power consumption. In addition, the required feedbacks to enable MIMO and MU-MIMO operation with so many spatial streams carry with them a significant airtime overhead that reduces the overall gains due to spatial streams growth.
  - –Station perspective: With the growing number of wireless interfaces in smartphones and PCs, it is very challenging to integrate all the required antennas into devices. Therefore, while some devices (gaming laptops, all-in-one PCs, etc.) may allow having a third or even a fourth Wi-Fi antenna, this will be a niche of the market, and therefore of less interest. Further, standards already support designing Wi-Fi stations with more than two antennas, and therefore there is nothing that needs to be done from a standard development perspective.
  - –Regardless of the target device (AP or STA), since antenna size is linear with signal wavelength, smaller antennas can be used for transmissions in higher frequency bands. This can be used to increase the spatial resolution of transmissions and/or allow for the use of more spatial streams — opening the way for further data rate growth.

Another source of increased latency is interference from other networks and transmitters. Within the context of Wi-Fi, these could be OBSS transmitters, or devices participating in direct client-to-client communications. Wi-Fi 6 technology started to relate to OBSS interference and coexistence by introducing spatial reuse (SR) features, the majority of which are not part of the Wi-Fi 6 certification. Wi-Fi 6 failed to define a practical scheme for coexistence between infrastructure and peer-to-peer (P2P) Wi-Fi communications, the latter of which could happen over Wi-Fi Direct (WFD), TDLS, Wi-Fi Aware, or proprietary P2P Wi-Fi protocols. One way to improve the coexistence of Wi-Fi infrastructure networks and P2P Wi-Fi could be by dedicating different channels to networking and P2P communications, or by allocating specific TWTs for P2P operation. While this could be done proprietarily, Wi-Fi 6 did not define mechanisms for negotiating these. It is further currently uncertain whether 802.11be Triggered Tx Op Sharing, another way to address BSS-P2P coexistence, will eventually be part of Wi-Fi 7 technology, or whether there will be need to rely on proprietary methods to address the need.

Technical Direction for Wi-Fi 8 and Beyond

Comparing the discussions that predated the 802.11be task group to the evolving Wi-Fi 7 standard, it is now clear that several major features, namely, MIMO enhancements, multi-AP coordination, and potentially TxOP sharing, are not expected to make it into the final 802.11be standard [13]. Further, learning from experience, not all features defined in the standard will be part of an eventual Wi-Fi Alliance certification, and at least some of the promise of Wi-Fi 7 will only be realized in Wi-Fi 8. Similarly, it is expected that some of the features and capabilities discussed for Wi-Fi 8 will only materialize (in standards or products) in following generations.

We therefore expect Wi-Fi 8 and future Wi-Fi standards to address the technical challenges and use cases discussed in the previous section. To achieve this, it is expected that future Wi-Fi standards in the sub 8 GHz band will focus on delivering increased network capacity, as well as better QoS and higher reliability. These will be achieved through a combination of different capabilities such as collaborative multi-device operation modes (e.g., multi-AP coordination), new transmission schemes (e.g., full duplex), and increased AP managed network operation, including support for network managed P2P operation.

Discussions on initial potential directions and use cases for Wi-Fi 8 technology have started in the 802.11 Wireless Next Generation standing committee (WNG SC) in 2022. Most of the use case discussions have related to Metaverse and XR usages, IIoT, improvements in P2P operations, and in overall networking performance. Contributions discuss the following vectors [14]:

- Increasing peak data rates (×2 or more)
- Reducing latency to below 1 ms and potentially enabling deterministic transmissions
- Further increasing Wi-Fi connections' reliability and overall system efficiency
- Improving network manageability
- Reducing device-level power consumption
- Improving mobility and mesh support
- Improving coordination of BSS and P2P transmissions and enhancing performance in spectrum congested environments

A radically different direction proposed was to enable network scale improvements through machine learning (ML). Machine learning is already utilized to simplify network management and optimization in 5G networks, where today's solutions rely on most networks being "single vendor" networks. Standardizing Wi-Fi for ML could potentially allow ML based network optimizations even for multi-vendor heterogenous networks. It is yet uncertain whether this approach would be addressed through the emerging Wi-Fi 8 project in IEEE or through a dedicated TG.

Some of the capabilities and directions discussed have already been discussed by IEEE 802.11 in studies leading up to the formation of the 802.11be task group (e.g., in-band full duplex). Others have been discussed within the 802.11be task group (e.g., multi-AP coordination, joint transmission, and constellation shaping). Yet another group will be a new set of capabilities to first be discussed towards Wi-Fi 8.

Another potential direction for Wi-Fi 8 is to define use-case-specific modes of operation that, while ensuring the capability to coexist with legacy devices, will be optimized for managed environments, where it is often feasible to meet the stringent QoS requirements of XR, TSN, and IIoT.

Regardless of the eventual directions taken in existing Wi-Fi bands of operation, several WNG contributions mentioned the potential for Wi-Fi expansion into mmWave bands, and specifically into the unlicensed 60 GHz band. This band has up to 14 GHz of unlicensed spectrum in the 57–71 GHz range, more than 10× the available

unlicensed spectrum below 10 GHz. It should be noted that 802.11 already defines modes of operation in this band (through the .11ad/ay amendments), but with limited market success. The underlying reasons for this limited success could have been the market introduction timing and/or the lack of a "killer use case." Another factor could be that the significant implementation differences between sub 8 GHz Wi-Fi and the 802.11ad/ay-based solution was a significant enough barrier for the integration of 802.11ad/ay solutions with the then incumbent 802.11ac-based Wi-Fi in devices. This resulted in increased product costs and reduced adoption rates, and eventually with several companies stopping their investment in the technology.

Therefore, we believe that a new approach is required to succeed in enabling the band for Wi-Fi operation, where a future 802.11 millimeter-wave (mmWave) program would make "maximizing synergies with existing sub-8 GHz Wi-Fi one of its the key tenants when reconsidering use of the 60 GHz band within a potential Wi-Fi 8 program in IEEE. One way to achieve this is to adapt one of the existing sub-8 GHz PPDU formats to mmWave operation by, for example, clock scaling. With the 60 GHz band (and related features such as multipath spread, coherency time, etc.,) being ~ ×10 of the 5–7 GHz band, a clock factor of 10 (or 8) could potentially make sense. Using this approach might allow using a similar architectural design for sub-8 GHz and mmWave Wi-Fi and reduce the development and support investment. Some adaptations will of course need to be made in relation to the unique band characteristics, such as analog beamforming and ease of initial discovery in conjunction with it — potentially learning from 11ad/ay.

It should be noted that in addition to networking and other communication usages, the mmWave band brings significant benefits to sensing use cases. Indeed the 802.11bf TG is in the process of defining sensing dedicated waveforms in the 60 GHz band. In case Wi-Fi 8 does go with the proposed alignment to sub-8 GHz Wi-Fi PPDU formats, and in order to align to a single PPDU format in the band, it might be required to review any work in .11bf and see how to best enable mmWave Wi-Fi sensing as part of Wi-Fi 8 (or derived) work.

Conclusion

Wi-Fi has been one of the most successful wireless technologies to date and has delivered multiple innovations throughout the years, with data rates growing from a few megabits per second to multi-gigabit-per-second links using capabilities such as OFDM, single-user and multi-user MIMO, and varying channel sizes. In addition to data rate growth, the introduction of OFDMA and trigger-based access significantly improved access efficiency and allowed increased device density and reduced latency in heavily used networks.

As we begin the journey toward Wi-Fi 8 and beyond, we expect that the pace of Wi-Fi innovation will continue, and that Wi-Fi will continue to be the main method to access the Internet and emerging metaverse, while at the same time proliferating to serve private networks and emerging use cases around AR/VR and Industrial IoT.

References

[1] "The Economic Value of Wi-Fi®: A Global View (2021–2025)," Wi-Fi Alliance, Sept. 3, 2021; https://www.wi-fi.org/file/detail-global-economic-value-of-wi-fi-2021-2025, accessed Feb. 1, 2022.

[2] E. Khorov, I. Levitsky, and I. F. Akyildiz, "Current Status and Directions of IEEE 802.11be, the Future Wi-Fi 7," *IEEE Access*, vol. 8, 2020, pp. 88,664–88. DOI: 10.1109/ACCESS.2020.2993448.

[3] Intel, "Intel Launches Intel® Centrino™ Mobile Technology," Mar. 12, 2003; https://www.intel.com/pressroom/archive/releases/2003/20030312comp.htm, accessed Feb. 1, 2022.

[4] Wi-Fi Alliance, "Wi-Fi Alliance® Introduces Wi-Fi 6," Oct. 3, 2018; https://www.wi-fi.org/news-events/newsroom/wi-fi-alliance-introduces-wi-fi-6, accessed Feb. 1, 2022.

[5] Wi-Fi Alliance, "Wi-Fi CERTIFIED™ AC Takes Wi-Fi® Performance to New Heights," June 19, 2013; https://www.wi-fi.org/news-events/newsroom/wi-fi-certified-ac-takes-wi-fi-performance-to-new-heights, accessed Feb. 1, 2022.

[6] FCC, "FCC Opens 6 GHz Band to Wi-Fi and Other Unlicensed Uses," Apr. 24, 2020; https://www.fcc.gov/document/fcc-opens-6-ghz-band-wi-fi-and-other-unlicensed-uses-0, accessed Feb. 1, 2022.

[7] "6GHz for License-Exempt Access"; https://6ghz.info, accessed July 1, 2022.

[8] Wi-Fi Alliance, "Wi-Fi CERTIFIED 6™ Release 2 Adds New Features for Advanced Wi-Fi® Applications," Jan. 5, 2022; https://www.wi-fi.org/news-events/newsroom/wi-fi-certified-6-release-2-adds-new-features-for-advanced-wi-fi-applications, accessed Feb. 1, 2022.

[9] CNET, "Wi-Fi 6 is Barely Here, but Wi-Fi 7 Is Already on the Way," Sept. 3, 2019; https://www.cnet.com/tech/mobile/wi-fi-6-is-barely-here-but-wi-fi-7-is-already-on-the-way/, accessed Feb. 1, 2022.

[10] M. Park *et al.*, "Enhanced Multi-Link Single Radio Operation," Apr. 21, 2020; IEEE 802.11-20/0562r1.

[11] C. O. Jaynes *et al.*, "The Metaverse: A Networked Collection of Inexpensive, Self-Configuring, Immersive Environments," *Proc. Wksp. Virtual Environments*, Zurich, Switzerland, 2003, pp. 115–24. DOI: 10.2312/EGVE/IPT_EGVE2003/115-124.

[12] R. d. Vegt, "Reduced Latency Benefits of Wi-Fi 6 OFDMA," Mar. 22, 2021; https://www.wi-fi.org/beacon/rolf-de-vegt/reduced-latency-benefits-of-wi-fi-6-ofdma, accessed Feb. 1, 2022.

[13] L. Cariou *et al.*, "Discussion on 11be R2 Scope," Nov. 8, 2021; IEEE 802.11-21/1598r1.

[14] Gan *et al.*, "Looking Ahead to Next Generation: Follow-Up," Mar. 8, 2022; IEEE 802.11-22/0458r1.

[15] Wi-Fi QoS Management; https://www.wi-fi.org/discover-wi-fi/wi-fi-qos-management, accessed July 1, 2022

Biographies

Ehud Reshef [SM] (ehud.reshef@intel.com) received his B.Sc. in electrical engineering from the Technion, Haifa, Israel, in 1991, and his M.B.A and M.Sc. in electrical engineering from the Tel Aviv University, Israel, in 1998 and 2006, respectively. He is currently a principal engineer with the wireless communication solutions group of Intel Corporation, where he is responsible for Wi-Fi product innovation and technology strategy. He has been a wireless innovator for more than 20 years and holds over 37 U.S. patents.

Carlos Cordeiro [F] (carlos.cordeiro@intel.com) is an Intel Fellow serving as the wireless CTO at Intel's client group. He is responsible for Intel's next generation wireless connectivity technology strategy, standards, ecosystem engagements, and regulatory strategy. He serves as the Chair of the Wi-Fi Alliance Board of Directors and the Associate Editor-in-Chief of *IEEE Communications Standards Magazine*, and has served as an editor of various journals including *IEEE Transactions on Mobile Computing* and *IEEE Transactions on Wireless Communications*. He is the co-author of two textbooks on wireless communications, has published over 120 papers, and holds over 400 patents.

Machine Learning and Analytical Power Consumption Models for 5G Base Stations

Nicola Piovesan, David López-Pérez, Antonio De Domenico, Xinli Geng, Harvey Bao, and Mérouane Debbah

The authors exploit the knowledge gathered by this framework to derive a realistic and analytically tractable power consumption model, which can help drive both theoretical analyses as well as feature standardization, development, and optimization frameworks.

Abstract

The energy consumption of the fifth generation (5G) of mobile networks is one of the major concerns of the telecom industry. However, there is not currently an accurate and tractable approach to evaluate 5G base stations' (BSs') power consumption. In this article, we propose a novel model for a realistic characterization of the power consumption of 5G multi-carrier BSs, which builds on a large data collection campaign. At first, we define a machine learning architecture that allows modeling multiple 5G BS products. Then we exploit the knowledge gathered by this framework to derive a realistic and analytically tractable power consumption model, which can help driving both theoretical analyses as well as feature standardization, development, and optimization frameworks. Notably, we demonstrate that this model has high precision, and it is able to capture the benefits of energy saving mechanisms. We believe this analytical model represents a fundamental tool for understanding 5G BSs' power consumption and accurately optimizing the network energy efficiency.

Introduction

The fifth generation (5G) of radio technology has brought about new services, technologies, and networking paradigms, with the corresponding societal benefits. However, the energy consumption of the new 5G network deployments is concerning. Deployed 5G networks have been estimated to be about 4× more energy-efficient than 4G ones. Nonetheless, their energy consumption is around 3× larger, due to the larger number of cells needed to provide the same coverage at higher frequencies, and the increased processing required by its wider bandwidths and more antennas [1]. To put this number into context, it should be noted that, on average, the network operational expenditure (OPEX) already accounts for around 25 percent of the total operator's cost, and 90 percent of it is spent on large energy bills [2]. Notably, most of this energy — more than 70 percent — has been estimated to be consumed by the radio access network (RAN), and in more detail, by the base stations (BSs), while data centers and fiber transport only account for a smaller share [3, 4].

To decrease the RAN energy consumption, Third Generation Partnership Project (3GPP) New Radio (NR) Release 15 specified intra-NR network energy saving solutions, similar to those developed for 3GPP Long Term Evolution (LTE), such as autonomous cell switch-off/re-activation capabilities for capacity booster cells via X_n/X_2 interfaces. Moreover, 3GPP NR Release 17 has recently specified inter-system network energy saving solutions, and is currently taking network energy saving as an artificial intelligence use case. However, data gathered about the benefits brought by 3GPP LTE and NR energy saving solutions have shown that they are not enough to fundamentally address the energy consumption challenge [5].

To continue tackling this challenge, 3GPP NR Release 18 has recently approved a new study item, "Study on NR Network Energy Saving Enhancements," which attempts to develop a set of more flexible and dynamic network energy saving solutions [5]. In more detail, the main objectives of this study item are:

- Identify new energy saving scenarios beyond that of the capacity booster cell (e.g., compensation cells).
- Study enhancements to allow faster adaptation of networking resources to traffic needs through:
  - User equipment (UE) assistance information reports
  - BS information exchange to share traffic predictions and support both beam-level operation and transmit power adjustment coordination
  - Downlink (DL)/uplink (UL) channel measurement enhancements

Importantly, to analyze the gains brought by such new schemes, there has been consensus on the need for new models to accurately estimate the 5G network power consumption. 3GPP NR Release 16 defined a power consumption model for 5G UEs [6]. However, there is no 5G network counterpart. Ongoing 3GPP discussions have suggested that such a new 5G network power consumption model should be a function of the number of BSs in the area of study, their frequency of operation, bandwidth, transmit power, number of transceivers, signaling configuration, physical resource blocks (PRBs) load, multiple-input multiple-output (MIMO) layers usage, as well as energy saving functionalities and their related sleep states and transition times.

To fill this gap, in this article, we introduce a new power consumption model for 5G active antenna units (AAUs), the highest power con-

[1] In 5G terminology, a massive MIMO BS is divided into three parts: the centralized unit, the distributed unit, and the AAU.

Digital Object Identifier: 10.1109/MCOM.001.2200023

The authors are with Huawei Technologies, France.

0163-6804/22/$25.00 © 2022 IEEE

suming component of a BS[1] and in turn of a mobile network. In particular, we present an analytically tractable model, which builds on a large data collection campaign and our machine learning (ML)-based analysis. The proposed model is realistic, as it is characterized by high precision, and generalizes well to a high number of 5G AAU types/products. For example, it accounts for multi-carrier AAUs embracing the widely used multi-carrier power amplifier (MCPA) technology [7].[2] This allows sharing some of the PA hardware among multiple carriers managed by an AAU, thus reducing its power consumption. Moreover, our model also captures the benefits brought by complex, standardized shutdown schemes (i.e., carrier shutdown, channel shutdown, symbol shutdown, and deep dormancy) [4] when operating in the field.

About the methodology adopted in this article, it should be highlighted that the parameters of the proposed analytical model are derived for a selected AAU product by using data collected from a real network deployment. Unfortunately, however, it is generally not possible to obtain exhaustive data for all possible input configurations for all AAU products deployed in real networks. Importantly, the inaccessibility of AAU measurements of power consumption under some conditions may prevent the derivation of the analytical model parameters. Therefore, we implement a methodology in which an ML framework is designed and trained to gather knowledge from many different types of AAUs with different hardware configurations. Notably, this modeling approach allows taking advantage of the ML generalization properties, generating synthetic data covering scenarios that may not be directly observable in the collected data but are needed to derive the proposed analytical model.

Related Works on BS Power Model

It has been reported that 73 percent of the total network energy is consumed by the BSs [3], where the power amplifier, the transceivers, and the cables consume about 65 percent of the total BS energy [4]. Therefore, significant attention has been directed toward reducing the energy consumed by the BSs during recent years, and various BS power consumption models have been proposed and investigated as a result.

The work in [8] proposed one of the most widely used BS power consumption models in the literature. In particular, this model explicitly shows the linear relationship between the BS power consumption and its transmit power. Embracing the model in [8], the work in [9] proposed an extension, which additionally supports massive multiple-input multipleoutputs (mMIMOs) and energy saving capabilities, considering different sleep depths and transition times between different energy states. However, multi-carrier and/or carrier aggregation (CA) capabilities were not considered, and massive MIMO (mMIMO) power consumption estimations seem inaccurate [10], with an optimistic 40.5 W per BS.

With regard to mMIMO, the work in [11] extended the BS power consumption model in [8], considering a linear increase of the power consumption with the number of mMIMO transceivers. More advanced works followed in this area, highlighting the importance of taking the impact of multi-UE scheduling and other mMIMO BS components into account in the modeling of 5G BS power consumption, such as power amplifiers, transceivers, analog filters, and oscillators. Specifically, the cornerstone research in [12] provided a more complete model, which considers the mMIMO BS architecture, both DL and UL communications, as well as the number of UEs multiplexed per PRB, and a large number of mMIMO BS components.

When modeling the power consumption of a system using multiple carriers and/or CA, it is also necessary to take into account how the power consumption scales with the number of component carriers (CCs) managed by the BS. The work in [13] captured this relationship using a linear model, but the literature is sparse in this area.

The work in [14] further combined and extended the linear version of the above presented works, jointly considering mMIMO and multi-carrier capabilities features, such as CA and its different aggregation capabilities: intra-band contiguous, intra-band non-contiguous, and inter-band.

5G AAU Architecture Model

Although various aspects related to the power consumption of a 5G BS have been considered in the research presented earlier, the true complexity of a 5G multi-carrier mMIMO AAU, where a single power amplifier may accommodate multiple carriers using MCPA technology, is not embraced in any of them. Our article fills this gap by defining a general and practical AAU architecture, and providing first the corresponding data-driven power consumption model, and then an analytical formulation fitted with realistic values for a particular AAU type.

In more detail, in our AAU architecture, we assume that:

- The AAU has a multi-carrier structure, and uses MCPA technology.
- The AAU manages C carriers — or CCs, using CA terminology — deployed in T different frequency bands.
- The AAU comprises T transceivers, each operating at a different frequency band, and M MCPAs, one for each antenna port.
- A transceiver includes M radio frequency (RF) chains, one per antenna port, which comprehend a cascade of hardware components for analog signal processing, such as filters and digital-to-analog converters.
- Antenna elements are assumed to be passive. For example, one wideband panel or T antenna panels may be used per AAU.
- Deep dormancy, carrier shutdown, channel shutdown, and symbol shutdown are implemented, each switching off distinct components of the AAU.

Figure 1 shows the AAU architecture and its main power consumption components. In more detail, the overall AAU consumed power includes:

- The baseline power consumption, P_0, which accounts for part of the AAU circuitry that is always active (e.g., circuitry used to control the AAU activation/deactivation)
- The power consumption, P_{BB}, required for the baseband processing performed at the AAU

About the methodology adopted in this article, it should be highlighted that the parameters of the proposed analytical model are derived for a selected AAU product by using data collected from a real network deployment. Unfortunately, however, it is generally not possible to obtain exhaustive data for all possible input configurations for all AAU products deployed in real networks.

[2] An MCPA operates, in contrast to a single-carrier power amplifier (PA), on multiple carriers as input, and provides a single amplified output.

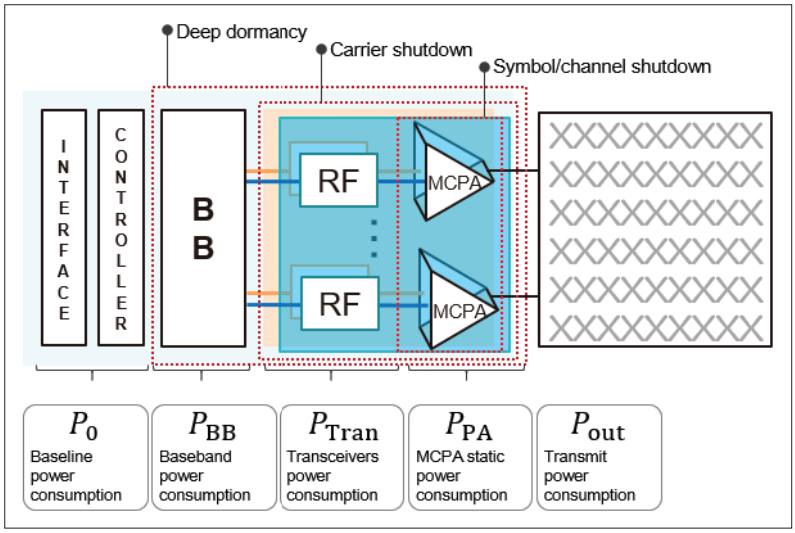


FIGURE 1. AAU with MCPAs handling 2 CCs in 2 different bands, which transmits over the same wideband antenna panel.

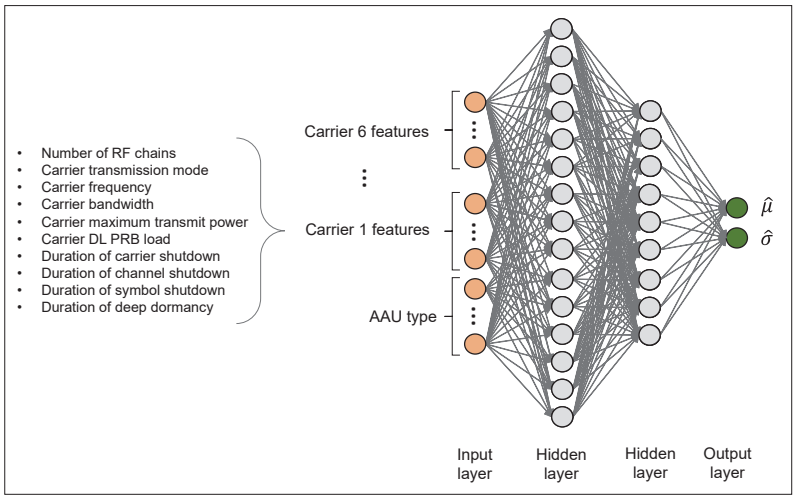


FIGURE 2. Architecture of the designed ANN.

- The power consumed by the T transceivers in the AAU, P_{Tran}
- The static power consumed by the MCPAs, P_{PA}
- The power consumed to generate the transmit power required to transmit the data over the C CCs, P_{out}

As described in [7], it should be highlighted that the implementation of MCPAs results in increased energy efficiency with respect to single-carrier power amplifiers. In more detail, by integrating multiple carriers together, the total transmit power managed becomes greater, thus enabling MCPAs to operate at higher efficiency areas. Moreover, the static power consumption of the MCPAs increases sublinearly with respect to the number of carriers, as some of the signal processing components can be shared among them. However, it is worth highlighting that the implementation of MCPAs entails increased complexity in the management of the network energy saving, and thus in the estimation of power consumption. In fact, contrary to what is commonly considered by simplistic models, the deactivation of just one carrier may not bring the expected energy savings if the MCPAs need to remain active to operate other co-deployed active carriers.

Artificial Neural Network Model

In this section, we describe the measurements gathered during our data collection campaign. Moreover, we provide a detailed description of the implemented artificial neural network (ANN) architecture for modeling and estimating power consumption, as well as an analysis of its accuracy. Note that ANNs were selected after evaluating and comparing their performance with those of other ML methods. The better performance of ANNs emanates due to their better capabilities to deal with the available tabular data and superior generalization properties.

Dataset

We collected hourly measurements for 12 days from a real deployment with 7760 5G AAUs in China, comprising 25 different types of AAU from a single vendor. Note that such data contains sensitive information regarding proprietary product hardware specifications, which cannot be made publicly available. The gathered information contains 150 different features, which can be divided into four main categories:

- *Engineering parameters*: Information related to the configuration of each AAU (e.g., AAU type, number of RF chains, numbers of supported and configured carriers)
- *Traffic statistics*: Information on the serviced traffic (e.g., average number of active UEs per transmission time interval, number of used PRBs, traffic volume)
- *Energy saving statistics*: Information on the activated energy saving modes (e.g., duration of the carrier, channel and symbol shutdown as well as dormancy activation)
- *Power consumption statistics*: Information on the power consumed by the AAUs

Inputs of the Model

Feature importance analysis was performed to identify the most relevant input features in the available dataset. Such features are the type/model of AAU, together with the key characteristics of the configured carriers. To give an example, such key characteristics comprehend, among others, frequency- and power-related engineering parameters, such as the carrier frequency, bandwidth and transmit power, the DL PRB load, and the amount of time for which each energy saving mode is activated. See Fig. 2 for a detailed description of all the selected input features. Note that the identified features are fundamental parameters, which are available in the products of any vendor. Moreover, feature importance analysis can extend the inputs of our ANN model to consider proprietary and not standardized energy saving schemes. After selection, each of the input features was pre-processed and then represented by one or more neurons at the input layer of the ANN. The numerical features were normalized before being input to the model, whereas the categorical ones were input by using one-hot-encoding.

Since a 5G AAU can operate multiple carriers through an MCPA, to make our ANN model the most general and flexible, the input layer takes input from the maximum number of carriers that can be managed by the most capa-

ble AAU in the dataset. When no carriers are deployed in an AAU, the input neurons related to the deployed carriers are set to zero. This approach allows our ANN model to be implemented with a fixed number of input neurons, and thus construct a single model for all possible AAU types and carrier configurations with minimal accuracy loss. The maximum accuracy loss observed when comparing this single model approach with respect to training different models for the different AAU types and carrier configurations is 1.86 percent.

Outputs of the Model

Different power consumption values are observed in the data for the same input feature values due to missing input features and/or errors in the measurements or in the collection/processing of the data. To embrace such noise, we define the measured power consumption for a given input configuration plus a noise originating from the mentioned errors. The analysis of the available data highlighted that such noise is normally distributed. It thus follows that the measured power consumption is normally distributed.

With this in mind, the goal of our ANN model is to produce, for a given input configuration, an estimate of the mean and standard deviation of the power consumption distribution. This allows having an evaluation of the confidence interval for each of the performed power consumption estimations during training and testing and in turn increasing the reliability of the whole estimation process.

Model Architecture

We consider the multilayer perceptron as basic architecture for the ANN, consisting of multiple fully connected layers of neurons [15]. The overall architecture of the proposed ANN model is also depicted in Fig. 2.

In our specific scenario, we collected data for 25 different AAU types, where the most capable AAU supports up to 6 CCs. The input layer is thus composed of 85 neurons, and it is followed by two hidden layers, which comprise 100 and 50 neurons, respectively. These dimensions were chosen after an optimization process targeted at maximizing the model accuracy. Finally, the output layer is composed of two neurons, which capture the mean and the standard deviation of the power consumption, as explained earlier.

Training of the Model

The model is trained with the objective of reducing both the prediction error and its uncertainty. In particular, the training is considered successful if the distribution output by the model for a given input matches the distribution of the power measurements in the data.

In terms of data management, we split the available dataset related to 7760 AAUs into two parts: a training set and a testing set. The training set contains data collected for 10 days, whereas the testing set contains the data collected for the 2 remaining days. The model training was performed by adopting the Adam version of the gradient descent algorithm [15], and required 75 minutes to perform 1086 iterations.

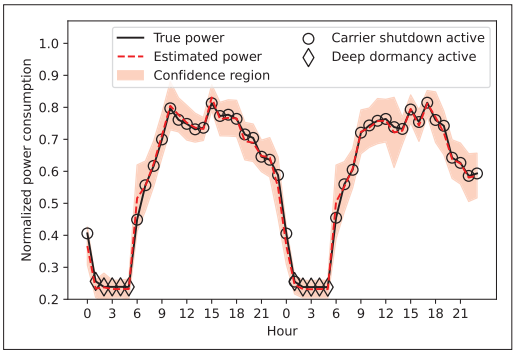


FIGURE 3. Hourly real and estimated normalized power consumption for an AAU doing carrier shutdown and deep dormancy.

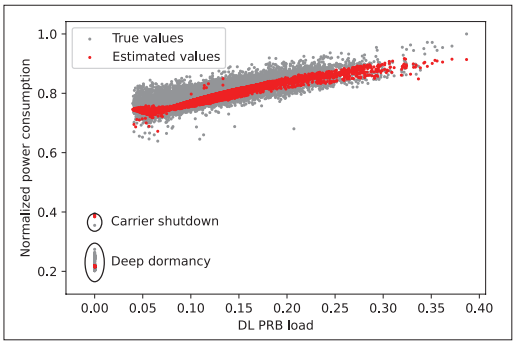


FIGURE 4. Normalized power consumption, estimated by the ANN model, vs. AAU DL PRB load for the selected AAU.

Model Performance Evaluation

To assess the performance of the proposed ANN model, we compared the estimated power consumption during the testing phase with the real measurements available in the data. Overall, the model achieved a root mean square error (RMSE) of 25.02 W, a mean absolute error (MAE) of 12.21 W, and a remarkably low mean absolute percentage error (MAPE) of 6.55 percent when estimating the power consumed by each AAU in each hour of the testing period.

To highlight the ability of the model to accurately estimate the power consumption when dynamic energy saving algorithms are activated, Fig. 3 shows an example of the real and estimated power consumption of a particular AAU, which supported up to six carriers and intensively used energy saving features during the two testing days. The confidence region is also reported, representing the interval in which the true power consumption is expected to fall with a 0.95 probability. Note that we normalized the power consumption for privacy reasons. From this figure, it can be observed that the deep dormancy feature is activated during night hours (i.e., from 1 a.m. to 6 a.m.), while the carrier shutdown algorithm is activated and intensively used during the rest of the day. Note that an AAU can only shut down a carrier when its shutdown entry conditions are met, and that such conditions mostly depend on traffic load and are independently checked per carrier on a less than a minute basis. Overall, even in these highly dynamic activation/deactivation conditions, the proposed model is able to estimate the power consumption of this AAU with high accuracy (i.e., RMSE 14.43 W, MAE 9.5 W, MAPE 2.5 percent).

The model is trained with the objective of reducing both the prediction error and its uncertainty. In particular, the training is considered successful if the distribution output by the model for a given input matches the distribution of the power measurements in the data.

To analyze the proposed analytical model performance, we have fitted its parameters, for the popular AAU type introduced earlier, by using power consumption estimations performed through our ANN model.

To highlight the capability of the model to perform in a variety of deployment environments, Fig. 4 also shows the real and estimated power consumption, not of a single AAU as in Fig. 3, but for a popular AAU type also supporting up to six carriers, which appears often in our dataset in different scenarios and city areas, with respect to the DL PRB load. For the sake of clarity, we would like to highlight here that the spread of the real and estimated values observed over the y-axis in this figure is motivated, in addition to the noises introduced earlier, by the presence of multiple carriers deployed within this AAU type, which are generally configured with different maximum transmit powers. As a result, there is not a biunivocal relation between the DL PRB load and the total transmit power (and thus with the power consumption). From this figure, it can be seen that this AAU type achieves a 47 and 70 percent reduction in power consumption when doing carrier shutdown and deep dormancy, respectively. Importantly, even if this AAU type was deployed in a heterogeneous set of scenarios, the proposed model is able to accurately estimate the power consumption (i.e., RMSE 18.25 W, MAE 14.48 W, MAPE 2.63 percent).

Analytical Model

Although accurate and general, the presented ML model lacks tractability to drive energy efficiency feature standardization, development, and/or optimization. To facilitate these tasks, in this section, based on the knowledge gathered from the previous ML model, we propose an analytically tractable 5G AAU power consumption model, which is easily interpretable and amicable to optimization.

Model Description

Our proposed 5G AAU power consumption model, which characterizes the relationships between the key characteristics that play a major role on 5G AAU power consumption, is mathematically formulated as

$$P_{\mathrm{AAU}} = P_0 + P_{\mathrm{BB}} + \underbrace{+\sum_{t=1}^{T} M_{\mathrm{av,t}} D_{\mathrm{Tran,t}}}_{P_{\mathrm{Tran}}} + \underbrace{M_{\mathrm{ac}} D_{\mathrm{PA}}}_{P_{\mathrm{PA}}} + \underbrace{\frac{1}{\eta}\sum_{c=1}^{C} P_{\mathrm{TX},c}}_{P_{\mathrm{out}}}. \quad (1)$$

In more detail, the power, P_{Tran}, consumed by the tth transceiver in the AAU is the product of the number of available RF chains, M_{av}, and the power consumed by each RF chain, $D_{Tran,t}$. The static power, P_{PA}, consumed by the MCPAs is the product of the number of active RF chains, M_{ac}, and the static power consumed by each MCPA, D_{PA}. Recall that there is an MCPA for each RF chain spanning over the managed carriers. Finally, the power, P_{out}, consumed to generate the transmit power required to transmit the data over the cth CC is equal to the ratio of the transmit power in use at the CC, $P_{\mathrm{TX,c}}$, to the efficiency of the MCPAs and antennas, η, where the transmit power in use usually linearly increases with the number of PRBs utilized.

When symbol shutdown is activated, the AAU switches off the MCPAs, and its power consumption is reduced to the sum of the baseline power consumption, P_0, the baseband processing power consumption, P_{BB}, and the power consumed by the transceivers, P_{Tran}, as they are not deactivated.

When channel shutdown is active, the AAU reduces power consumption by limiting the multiplexing and beamforming capabilities of the cell (i.e., by limiting the number of active MCPAs). This is realized in our model by decreasing the value of the variable, M_{ac} (e.g., from 64 to 32 or 16).

When carrier shutdown is activated, the MCPAs and the transceivers are switched off. Therefore, the power consumption is further reduced to the sum of the baseline power consumption, P_0, and the baseband processing power consumption, P_{BB}. Finally, when deep dormancy is activated, the circuitry for baseband processing is switched off, and the AAU power consumption is further reduced to the baseline power consumption, P_0.

The proposed model has a number of benefits not captured by other models in the literature, which makes it a cornerstone for accurate 5G network energy efficiency standardization, development, and optimization:

- It allows capturing different multi-carrier architectures (i.e., intra-band contiguous, intra-band non-contiguous, and inter-band), where distinct carriers may or may not share the same transceiver.
- It characterizes realistic multi-carrier AAU products with MCPAs and their intricate shutdown functioning, where deactivating only a subset of the carriers in the AAU does not lead to large energy savings, since the MCPAs must continue operating to support the active carriers;
- It accounts for each of the state-of-the-art energy saving techniques (i.e., carrier shutdown, channel shutdown, symbol shutdown, and deep dormancy), and can easily be extended to more.

Model Fitting

To analyze the proposed analytical model performance, we have fitted its parameters for the popular AAU type introduced earlier by using power consumption estimations performed through our ANN model. Note that this approach to fitting, not based on the data but on the ANN model created through the data, allows us to exploit the generalization capabilities of our proposed ANN model, which can learn from other AAU types and perform accurate power consumption estimations for traffic conditions not observed in the data of this AAU type. In more detail, the analytical model parameters have been fitted on the generated data by iteratively solving a nonlinear least-squares regression problem. The normalized values for the fitted parameters are $P_0 = 0.22$, $P_{\mathrm{BB}} = 0.16$, $D_{\mathrm{Tran,1}} = 1.47 \cdot 10^{-3}$, $D_{\mathrm{PA}} = 3.81 \cdot 10^{-3}$, $\eta = 0.4$. For completeness, let us note that the AAU under study has $C = 2$ CCs with $M_{\mathrm{av,1}} = 64$ RF chains.

Model Performance Evaluation

Figure 5 shows the normalized power consumed by the selected AAU type for different values of the DL PRB load observed in the dataset and the values estimated by the fitted analytical model. The analytical model achieves remarkable performance with RMSE 19.96 W, MAE 16.50 W, and

MAPE 2.67 percent. This estimation accuracy is close to that achieved by the ANN model, highlighting that the most relevant inputs to power consumption have been captured, and the capability of the proposed analytical model to accurately model realistic AAUs, while considering the complex MCPA structure and the existence of different energy saving modes. A comparison of the accuracy performance reached by the ANN and analytical models is reported in Table 1.

From Fig. 5, it can also be observed that for this AAU type, the activation of symbol shutdown provides a 34 percent power consumption saving w.r.t to the power consumption at zero load, while that of carrier shutdown results in larger savings, 47 percent.

It should be noted, however, that the lower power consumption achieved by carrier shutdown comes at the expense of increased complexity in network management. Symbol shutdown operates locally — and usually opportunistically — in every cell at the timescale of hundreds of microseconds when no data needs to be transmitted, and thus it does not generally affect user performance. On the contrary, carrier shutdown strategies are adopted for longer time periods (from a few minutes to a few hours) and coordinated across the network, as their activation requires/implies the redefinition of the network coverage and redistribution of its traffic (i.e., user association). Due to its complexity, if the carrier shutdown feature is not appropriately optimized, energy savings may come at the expense of user experience. Even worse, if the optimization is performed with an inaccurate AAU power consumption model, the energy saving gains may not even be there.

To illustrate this point, we have estimated the power consumption of the selected AAU under the same conditions over the 24 hours of a day with a state-of-the-art power consumption model [12]. This model provided a 2.5× overestimation of the power consumption over the ground truth, as it was not able to capture the multi-carrier architecture and the accurate impact of energy saving methods. The error of our analytical model was less than 1 percent. This significant overestimation would lead to a suboptimal carrier shutdown configuration, hindering energy savings, and shows how state-of-the-art models may fail to drive network energy efficiency optimization. Instead, the better accuracy of our proposed model indicates that it may be a more viable tool to drive the optimization of greener 5G (and beyond) networks.

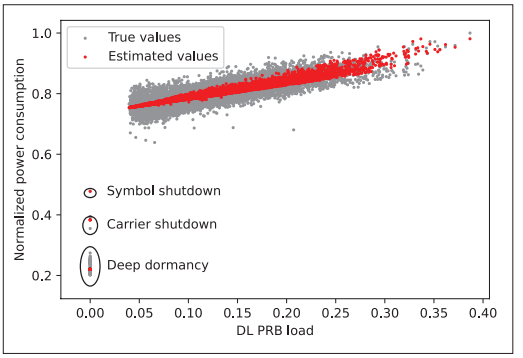


FIGURE 5. Normalized power consumption, estimated by the analytical model, vs. AAU DL PRB load for a popular AAU.

Metric	Analytical model	ML model	ML model gain
RMSE	19.96 W	18.25 W	8.6%
MAE	15.36 W	14.48 W	5.7%
MAPE	2.80%	2.63%	6.1%

TABLE 1. Comparison of accuracy performance achieved by the ANN model and the analytical model for a popular AAU type.

Conclusions

In this article, we present a novel power consumption model for realistic 5G AAUs, which builds on a large data collection campaign. At first, we propose an ANN architecture, which allows modeling multiple types of AAU and different configurations. The discussed results highlight that the designed ANN architecture is able to provide high accuracy. In a second stage, we exploit the knowledge gathered by the ANN method to derive a novel and realistic but analytically tractable 5G AAU power consumption model. We demonstrate that such an analytical model reaches accuracy close to that of the ML model for a widely used type of AAU. Notably, when compared to a state-of-the-art model under the same conditions, the proposed one is shown to be around 150 percent more accurate, as it is able to precisely capture the MCPA architecture and the benefits of shutdown approaches. Importantly, due to its fundamental nature, the proposed methodology can be adopted to model other types of AAU deployed in different multi-vendor networks.

We thus believe that this model is a valuable contribution to both industry and the research community working on wireless network energy efficiency and its optimization, and can be of use in the current 3GPP NR Release 18 work on network energy efficiency.

References

[1] Huawei Technologies Co., Ltd., "Green 5G: Building a Sustainable World," Tech. Rep., Aug. 2020; https://www.huawei.com0/en/ public-policy/green-5g-building-a-sustainable-world, accessed 19 Aug. 2022.

[2] GSMA, "5G Energy Efficiencies: Green is the New Black," Tech. Rep., Nov. 2020; https://data.gsmaintelligence.com/api-web/v2/ research-file-download?id=54165956&file=241120-5G-energy.pdf, accessed 19 Aug. 2022.

[3] —, "Going Green: Benchmarking the Energy Efficiency of Mobile," Tech. Rep., June 2021; https://data.gsmaintelligence.com/research/research/research-2021/going-green-benchmarking-the-energy-efficiency-of-mobile, , accessed 19 Aug. 2022.

[4] D. Lopez-Perez *et al.*, "A Survey on 5G Radio Access Network Energy Efficiency: Massive MIMO, Lean Carrier Design, Sleep Modes, and Machine Learning," *IEEE Commun. Surveys & Tutorials*, vol. 24, no. 1, 2022, pp. 653–97.

[5] China Telecom, "Draft SID on NR Network Energy Saving Enhancement (RWS-210152)," *3GPP TSG RAN Rel-18 Wksp.*, July 2021.

[6] T. Kim *et al.*, "Evolution of Power Saving Technologies for 5G New Radio," *IEEE Access*, vol. 8, 2020, pp. 198,912–24.

[7] S. Zhang *et al.*, "Dynamic Carrier to MCPA Allocation for Energy Efficient Communication: Convex Relaxation Versus Deep Learning," *IEEE Trans. Green Commun. and Networking*, vol. 3, no. 3, 2019 , pp. 628–40.

[8] G. Auer *et al.*, "How Much Energy Is Needed to Run a Wireless Network?," *IEEE Wireless Commun.*, vol. 18, no. 5, Oct. 2011, pp. 40–49.

[9] B. Debaillie, C. Desset, and F. Louagie, "A Flexible and Future-Proof Power Model for Cellular Base Stations," *IEEE*

VTC-Spring, 2015.
[10] S. Han, S. Bian *et al.*, "Energy-Efficient 5G for a Greener Future," *Nature Electronics*, vol. 3, no. 4, 2020 pp. 182–84.
[11] S. Tombaz *et al.*, "Energy Performance of 5G-NX Wireless Access Utilizing Massive Beamforming and an Ultra-Lean System Design," *IEEE GLOBECOM*, 2015.
[12] E. Björnson *et al.*, "Optimal Design of Energy-Efficient Multi-User MIMO Systems: Is Massive MIMO the Answer?" *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, 2015, pp. 3059–75.
[13] G. Yu *et al.*, "Joint Downlink and Uplink Resource Allocation for Energy-Efficient Carrier Aggregation," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, 2015, pp. 3207–18.
[14] D. López-Pérez *et al.*, "Energy Efficiency of Multi-Carrier Massive MIMO Networks: Massive MIMO Meets Carrier Aggregation," *IEEE GLOBECOM*, Dec. 2021.
[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016; http://www.deeplearningbook.org.

Biographies

Nicola Piovesan [M] (nicola.piovesan@huawei.com) is a senior researcher at Huawei Technologies, France. His research interests include energy sustainability, energy efficiency optimization, and ML in wireless communication systems.

David López-Perez [SM] is an expert and a technical lead at Huawei Technologies, France. His interests are in cellular and Wi-Fi networks, network performance analysis, network planning and optimization, as well as technology and feature development.

Antonio De Domenico [M] is a senior researcher at Huawei Technologies, France. His research interests include heterogeneous wireless networks, ML, and green communications.

Harvey Bao is a senior researcher at Huawei Technologies, France. His research interests include new technologies for wireless networks and AI-driven network modeling and optimization.

Xinli Geng is a principal engineer at Huawei Technologies, China. His current research interests include wireless networks optimization, green communications, data-driven network modeling, and AI-related technologies.

Mérouane Debbah [F] is chief research officer at the Technology Innovation Institute in Abu Dhabi. From 2014 to 2021, he was vice-president of the Huawei France Research Center, where he was jointly the director of the Mathematical and Algorithmic Sciences Lab as well as the director of the Lagrange Mathematical and Computing Research Center.

CALL FOR PAPERS

IEEE COMMUNICATIONS MAGAZINE

FEATURE TOPIC: THE INTERPLAY OF DIGITAL TWIN AND 6G WIRELESS NETWORKS

BACKGROUND

Over the last couple of decades, the paradigm of virtualization has been evolving from a virtualized local area network (LAN) and private networks to the solidification of network function virtualization (NFV) and network slicing principles. This advancement is driven by the edge computing and cloudification capabilities of current wireless network generations. With the growing demands of wireless networks, in terms of latency, reliability, and energy and spectral efficiency, and the emergence of sophisticated services with heavy distributed computing requirements, it is envisaged that the concept of network virtualization will be scaling up from the node and link levels to the network-wide level, setting the scene for a holistic network virtualization, from the core to the edge. Coupled with the pervasive utilization of artificial intelligence (AI) at all network levels, the Digital Twin (DT) paradigm has been recently deemed as a promising tool for network design, optimization, management, and recovery, in which the DT can be leveraged to realize the vision of sustainable, zero-touch 6G networks. The key principle of the DT paradigm is to create a virtual representation, not only for the physical elements, but also for the dynamics and functions of the network. According to its definition, the DT is envisioned to enable end-to-end digitization of wireless networks, with the aim to perform cost-effective, adaptive, efficient, and fast network-wide optimization of the available resources and infrastructure design. Furthermore, the DT allows the utilization of the digital realm with the aim to develop and test novel schemes and AI algorithms, that are capable of handling previously experienced critical situations or predicted scenarios based on the collected data at the cyber twin, and then to implement them at the physical twin once fully mature.

Despite its promising advantages, to reap the full potential of the DT technology in 6G networks, the cyber twin is envisaged to leverage AI algorithms, with novel data-driven paradigms, high performance computing, optimization theory, matching theory, as well as efficient cyber-physical interaction schemes, among others, to realize the necessary adaptation/reconfiguration at the physical twin with an imperceptible time-lag. In order to achieve the needed quality-of-service (QoS) for the successful implementation of a high-fidelity DT paradigm in 6G, a new level of stringent requirements pertaining to connectivity, reliability, latency, and data rate are imposed on future wireless generations. It is worth highlighting that the research on the interplay of the DT and 6G networks is still in its early stages, and the reported contributions in this field are very limited. While the field of DT-enabled wireless communication is relatively new, it has attracted the attention from the research community in a short span of time. Inspired by this, the aim of this Feature Topic is to attract novel research contributions on DT-empowered wireless networks, and to be a stepping-stone on advancing the research on DT for 6G. Such an attractive and new topic is expected to attract a large number of researchers, who are interested in exploring the advancements in DT-enabled 6G and its technical limitations and challenges.

The objective of this Feature Topic (FT) is to solicit research papers with original contributions that address the latest advances and challenges in DT-empowered wireless networks, paving the way for the efficient realization of pervasive intelligent and holistic network virtualization in future 6G networks. More specifically, this FT will bring together leading researchers from both industry and academia to present their views on this emerging research with respect to the fundamentals, core design aspects, applications, use-cases, and challenges of DT-empowered wireless networks.

Topics of interest include, but are not limited to:

- The interplay of machine learning and DT in 6G networks.
- Security and Privacy in DT-enabled 6G.
- The interplay of DT and the Metaverse.
- DT-based resource allocation in 6G networks.
- Latency minimization in DT-enabled wireless networks.
- DT for intelligent surface-assisted wireless networks.
- DT-assisted task offloading.
- DT for high-frequency wireless networks.
- Wireless edge-empowered DT.
- DT for industrial IoT.
- Novel AI Algorithms and Architectures for efficient DT.
- DT-enabled vehicular networks.
- DT for zero-touch networks.
- The interplay of DT and network slicing.
- DT for optical wireless communication.
- DT for satellite-enabled wireless communication.
- Sustainable wireless networks through DT.
- URLLC in DT-enabled wireless networks.
- Testbed designs and implementation of DT in wireless networks.
- Network simulations of DT-enabled 6G.

SUBMISSION GUIDELINES

Manuscripts should conform to the standard format as indicated in the Information for Authors section of the Manuscript Submission Guidelines. Please, check these guidelines carefully before submitting since submissions not complying with them will be administratively rejected without review.

All manuscripts to be considered for publication must be submitted by the deadline through Manuscript Central. Select the "FT-2222/The Interplay of Digital Twin and 6G Wireless Networks" topic from the drop-down menu of Topic/Series titles. Please observe the dates specified here below noting that there will be no extension of submission deadline.

IMPORTANT DATES

Manuscript Submission Deadline: 30 November 2022
Decision Notification: 15 March 2023
Final Manuscript Due: 31 March 2023
Tentative Publication Date: May 2023

GUEST EDITORS

Lina Bariah (Lead Guest Editor)
IEEE Tech. Innovation Institute, UAE
lina.bariah@ieee.org

Mérouane Debbah
Tech. Innovation Institute, UAE
CentraleSupelec, University Paris-Saclay, France
Merouane.Debbah@tii.ae

Hikmet Sari
Nanjing University of Posts and Telecommun. (NJUPT), China
hsari@ieee.org

Ejder Bastug
Nokia Bell Labs, France
ejder.bastug@nokia-bell-labs.com

Off-Network Communications for Future Railway Mobile Communication Systems: Challenges and Opportunities

Jiewen Hu, Gang Liu, Yongbo Li, Zheng Ma, Wei Wang, Chengchao Liang, F. Richard Yu, and Pingzhi Fan

The authors provide a comprehensive summary and analysis of off-network use cases in FRMCS. Then they give an overview of existing technologies (GSM-R, TETRA, DMR, LTE-V2X, and NR-V2X) that may support off-network communication.

Abstract

GSM-R is predicted to be obsoletd by 2030, and a suitable successor is needed. Defined by the International Union of Railways, the Future Railway Mobile Communication System (FRMCS) contains many future use cases with strict requirements. These use cases should ensure regular communication not only in network coverage but also in uncovered scenarios. There is still a lack of standards on off-network communication in FRMCS, so this article focuses on off-network communication and intends to provide reference and direction for standardization. We first provide a comprehensive summary and analysis of off-network use cases in FRMCS. Then we give an overview of existing technologies (GSM-R, TETRA, DMR, LTE-V2X, and NR-V2X) that may support off-network communication. In addition, we simulate and evaluate the performance of existing technologies. Simulation results show that it is possible to satisfy the off-network communication requirements in FRMCS with enhancements based on LTE-V2X or NR-V2X. Finally, we give some future research directions to provide insights for industry and academia.

Introduction

Over the past 40 years, the world has rapidly evolved from 1G to 5G. The emergence of new technologies means the obsolescence of old technologies, and the Global System for Mobile Communications-Railway (GSM-R), which is based on 2G, is no exception. The International Union of Railways (UIC) group started looking for a replacement of GSM-R in 2015. In 2018, they established a structured outline named Future Railway Mobile Communication System (FRMCS). FRMCS is intended to replace the soon-to-be obsolete GSM-R and is also oriented to future railway applications, such as automated train operation, remote control, and future fully self-driving trains [1]. This will require a more reliable and higher transmission rate communication system.

The issue of dedicated railway frequencies for FRMCS was raised as early as 2017, which boosted the development of FRMCS. In 2020, the European Conference of Postal and Telecommunications Administrations — Electronic Communications Committee (CEPT-ECC) decided to use the paired frequency bands 874.4–880.0 MHz and 919.4–925.0 MHz, and the unpaired frequency band 1900–1910 MHz for railway mobile radio. In 2021, China planned to use 2.1 GHz as the operating band for future 5G railway communication.

Based on the FRMCS user requirements specification [1] and use cases [2], the Third Generation Partnership Project (3GPP) studies and summarizes all the cases in FRMCS [3]. In addition, the 3GPP further elaborates the off-network use case scenarios and refines their performance requirements in [4]. Off-network communication refers to a direct communication mode between transmitter and receiver without passing the network.

The research on railway off-network communication is in the very preliminary stage. The study in [5] proposes a new train autonomous driving communication system based on 5G train-to-train (T2T) communication. The authors in [6] modified Long Term Evolution Vehicle-to-Everything (LTE-V2X) for T2T communication and proposed a train-centric communication-based train control (CBTC) system. Most of these studies focus on automatic train control (ATC), but there are many scenarios of off-network communication in FRMCS, which have not been comprehensively studied. The article [7] uses terrestrial trunked radio (TETRA) technology for T2T communication and performs channel measurements at 450 MHz. The authors of [8, 9] take practical measurements of intelligent transport systems (ITS-G5) in T2T scenarios and propose a geometry-based stochastic channel model (GSCM) for T2T communication in an open field environment. Although the existing works assume various existing technologies for railway off-network communication, there is still a lack of rigorous evaluation to judge whether they satisfy the requirements of off-network use cases in FRMCS.

This article first provides an overview of the off-network use cases in FRMCS, and introduces the existing technologies that may be used for future railway off-network communication, including GSM-R, TETRA, digital mobile radio (DMR), LTE-V2X, and New Radio-V2X (NR-V2X). The performance of existing technologies is compared through simulation. Based on the simulation results, it is suggested that the enhancements on

Jiewen Hu, Gang Liu (corresponding author), Yongbo Li, Zheng Ma, and Pingzhi Fan are with Southwest Jiaotong University, China; Wei Wang is with State Key Laboratory of Rail Transit Engineering Informatization (FSDI), China; Chengchao Liang is with Chongqing University of Posts and Telecommunications, China; F. Richard Yu is with Carleton University, Canada.

Digital Object Identifier: 10.1109/MCOM.001.2101081

0163-6804/22/$25.00 © 2022 IEEE

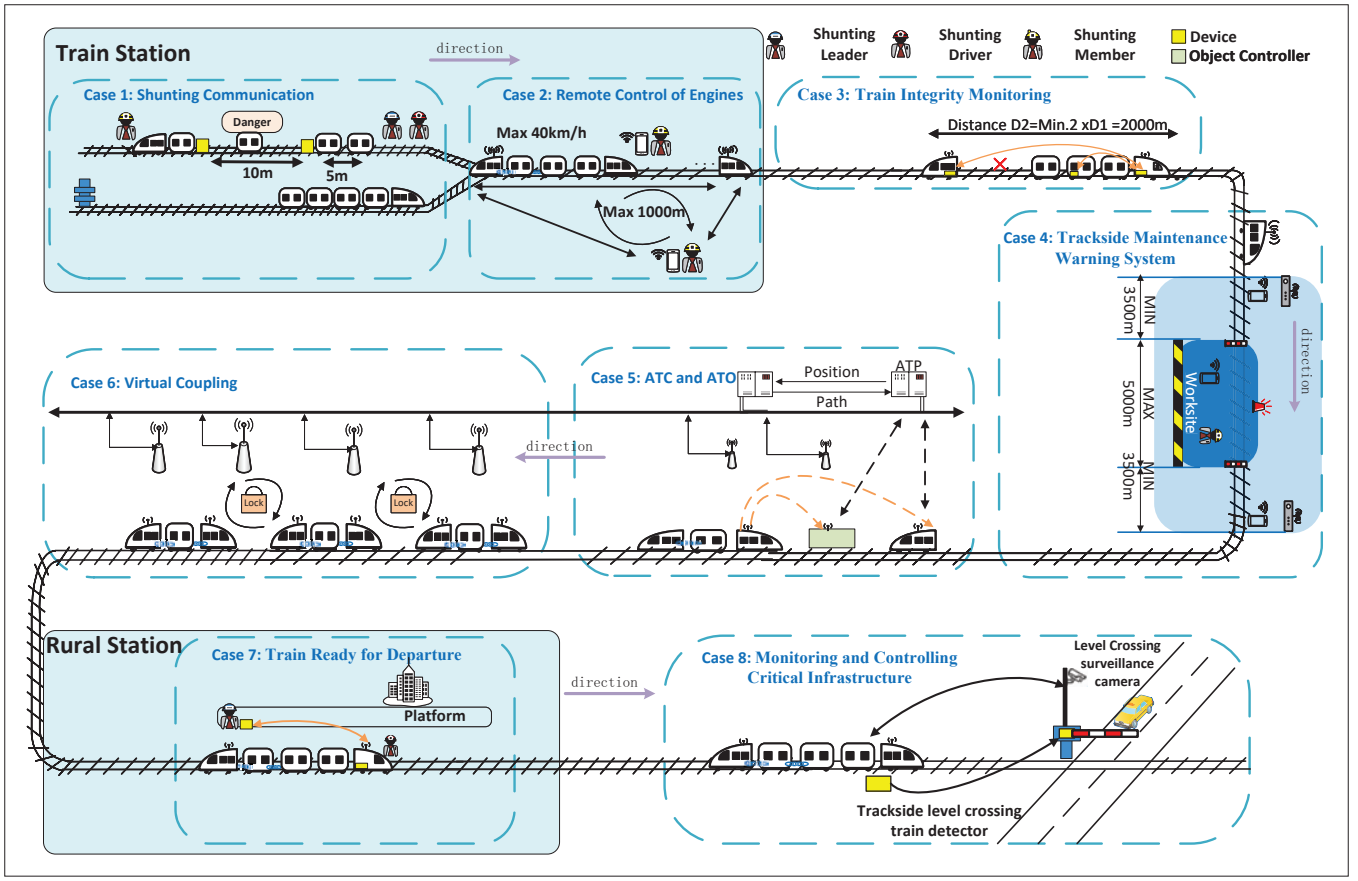


FIGURE 1. Use cases of off-network communication in FRMCS.

LTE-V2X or NR-V2X may satisfy the off-network communications requirements for FRMCS, which provides some potential directions and references for standardization. Finally, some possible research directions to improve transmission performance are given to provide insights for the future.

Use Cases and Requirements of Off-Network Communication for Railways

Figure 1 shows the use cases of off-network communication in FRMCS, and the performance requirements of off-network use cases are summarized in Table 1 [4].

Shunting Communication: Shunting movements include changing the locomotive of a train, coupling/uncoupling wagons, and changing the order in which wagons are arranged in a train. All shunting movements are done within a station and require cooperation between shunting members via radio, which contains data, voice, and video communications over the network or off-network.

Remote Control of Engines: The driver can remotely control the engine of the train via a ground-based system or an onboard system located at the other end of the engine. It is typically used for shunting operations in depots, shunting yards, or banking, and should support off-network communication due to the uncertainty of the network.

Train Integrity Monitoring: Accidental separation of train cabins is a hazardous event, and to avoid that situation, the driver can use this system to check the movement of the tail of the train while it is running. Since the track is not always covered by the network when the train is running, the system should also be capable of off-network communication.

Trackside Maintenance Warning System: To ensure the safety of the staff and the regular operation of the train, this system must promptly and accurately notify the staff of the upcoming train during the trackside maintenance period. Since the warning system is usually deployed temporarily and the network does not entirely cover the rail tracks, the system needs to have off-network communication capabilities.

ATC and Autonomous Train Operation (ATO): This system is expected to be used primarily for urban rail transportation such as subways. It allows trains to determine their movement autonomously based on direct communication between trains. For the safe operation of trains, even if the network is unavailable, the trains need to transmit speed, location, and other information in real time, so this system needs to have the ability to communicate off the network.

Virtual Coupling: Virtual coupling is when multiple trains in close distance move together as they are physically coupled. When the virtual coupling is formed, the distance between the trains is about 300 m (urban railway). This scenario is similar to vehicle platooning in V2X, where trains directly share control information (acceleration and braking, etc.) in real time. It can significantly reduce the distance between trains and increase the efficiency of railway transportation. Due to the high demand for latency, it is the only scenario that sets off-network communication as the default mode even when the network is available.

Train Ready for Departure: This scenario is

Scenario	End-to-end latency	Reliability	Data rate	Communication range
Shunting voice communication	≤ 100 ms	99.9999%	100 kb/s up to 300 kb/s	≤ 1.5 km
Shunting data communication	≤ 500 ms	99.9999%	10 kb/s up to 500 kb/s	≤ 1.5 km
Shunting video communication	≤ 100 ms	99.9%	10 Mb/s	≤ 1.5 km
Remote control of engines data communication	≤ 10 ms	99.9999%	100 kb/s up to 1 Mb/s	≤ 1 km
Remote control of engines video communication	≤ 100 ms	99.9%	10 Mb/s	≤ 1 km
Train integrity monitoring data communication	≤ 1 s	99.9%	10 kb/s up to 500 kb/s	≤ 2 km
Trackside maintenance warning system communication	≤ 500 ms	99.9999%	10 kb/s up to 500 kb/s	≥ 8.5 km
ATC and ATO	≤ 100 ms	99.99%	≤ 1 Mb/s	≤ 3 km
Virtual coupling critical data communication	≤ 100 ms	99.99%	≤ 1 Mb/s	≤ 3 km
Virtual coupling very critical data communication	≤ 10 ms	99.9999%	≤ 1 Mb/s	≤ 0.3 km
Train ready for departure data communication	≤ 500 ms	99.9%	10 kb/s up to 500 kb/s	≤ 1 km
Train ready for departure video communication	≤ 100 ms	99.9%	10 Mb/s	≤ 1 km
Monitoring and controlling critical infrastructure video communication	≤ 100 ms	99.9%	≤ 10 Mb/s	≤ 1 km

TABLE 1. Requirements of off-network communication use cases.

used to ensure that passengers can safely board the train and that the train can safely leave the platform. It consists primarily of driver-to-controller communication, conductor-to-driver communication, and platform camera video transmission. If the platform is in a remote area without a network, off-network communication is used.

Monitoring and Controlling Critical Infrastructure: This system is designed mainly for video communication and is usually used for intersection monitoring, train detection, signals, indicators, and so on. It should also be available in remote areas where there is no network coverage.

As shown in Table 1, most of the off-network communication use cases have strict performance requirements, which require a minimum latency of 10 ms, a maximum data rate of 10 Mb/s, and a maximum reliability requirement of 99.9999 percent. In addition, the communication range is typically within 3 km, except for the trackside maintenance warning system, which needs to achieve reliability of 99.9999 percent over a communication range of more than 8.5 km and is extremely difficult to achieve with the existing technologies.

In general, the provisions of these off-network use cases are full of challenges. We need to answer the question of whether any existing technology can satisfy these stringent requirements. If no, we should find a successor or enhance existing technologies.

Overview of Existing Technologies

This section provides a brief introduction to the existing technologies that may be used to support off-network communication. Among these technologies, GSM-R, TETRA, and DMR have been applied in some scenarios in the field of railway communication, while LTE-V2X and NR-V2X are technologies that can be used for off-network communication in the field of vehicular networking. The comparison of existing technologies is given in Table 2.

GSM-R [10]: GSM-R works near the 900 MHz frequency band, the access scheme is time-division multiple access (TDMA), and the channel bandwidth is 200 kHz. It can achieve data rates of 9.6 kb/s and 14.4 kb/s per channel using circuit switched data (CSD) mode and high-speed circuit-switched data (HSCSD) mode, respectively. In addition, it can achieve data rates of 21.4 kb/s per channel when the general packet radio service (GPRS) with packet switched data (PSD) is used.

GSM-R typically uses 1 W mean output power. The data TCHs use convolutional code as the channel coding scheme, and the modulation scheme of GSM-R is Gaussian minimum shift keying (GMSK) with parameter BT = 0.3 and modulation rate of 270.83 kb/s, and a Viterbi algorithm is used for demodulation.

TETRA [11]: TETRA operates at 300 MHz to 1 GHz, the access method is TDMA, and the channel bandwidth is 25 kHz. The peak rate of the data traffic channel is 7.2 kb/s. The average power of 0.56 W to 10 W is the common power range used by TETRA. The channel coding scheme for data traffic channels is rate-compatible punctured convolutional (RCPC) code. The modulation scheme is $\pi/4$-shifted differential quaternary phase shift keying ($\pi/4$-DQPSK), and the random access protocol used by TETRA is based on slotted Aloha.

DMR [12]: DMR operates in part of the frequency range from 30 MHz to 1 GHz, with a TDMA access scheme and a channel bandwidth of 12.5 kHz. The channel coding scheme of DMR is forward error correction (FEC) codes, including Golay code, Hamming code, block product turbo code (BPTC), and trellis code. The trellis code is used for data channels. The modulation scheme of DMR is 4-frequency shift keying (4FSK), and the average transmission power range is 1–40 W. The single-channel transmission rate can reach 4.8 kb/s.

LTE-V2X [13]: Since there are many similarities between railway communication and vehicle communication, V2X may be a good direction. LTE-V2X Mode 4 can directly communicate via sidelink without network support.

LTE-V2X operates in the 5.9 GHz (n47) band

Characteristics	GSM-R [10]	TETRA [11]	DMR [12]	LTE-V2X (sidelink) [13]	NR-V2X (sidelink) [14]
Access scheme	TDMA	TDMA	TDMA	SC-FDMA	OFDMA
Time slot duration	0.577 ms	14.167 ms	30 ms	0.5 ms	1/0.5/0.25/0.125/ 0.0625 ms
Transmission power	30 dBm	35 dBm	30-46 dBm	23 dBm	23 dBm
Frequency	Uplink: 876–915 MHz Downlink: 921–960 MHz	380-400 MHz 410–430 MHz 450–470 MHz 806–821 MHz 851–866 MHz	30 MHz–1 GHz	5855–5925 MHz (n47)	2570–2620 MHz (n38) 5855–5925 MHz (n47)
Channel bandwidth	200 kHz	25 kHz	12.5 kHz	1.4-20 MHz	5-40 MHz
Channel coding (data channel)	Convolutional Code	RCPC	Trellis Code	Turbo	LDPC
Modulation scheme	GMSK	$\pi/4$-DQPSK	4FSK	R14: QPSK 16-QAM R15: 64-QAM	QPSK 16-QAM 64-QAM 256-QAM
Peak transmission rate (single channel)	CSD: 9.6 kb/s HSCSD: 14.4 kb/s GPRS: 21.4 kb/s	7.2 kb/s	4.8 kb/s	30 Mb/s	1 Gb/s
Operation mode	Infrastructure-based Mode only	Infrastructure-based/ Direct Mode	Infrastructure-based/ Direct Mode	Infrastructure-based/ Direct Mode	Infrastructure-based/ Direct Mode
Access protocol (direct mode)	—	Slotted Aloha	Slotted Aloha	Sensing-based SPS	Sensing-based SPS
Multi-antenna supporting	N	N	N	Y	Y

TABLE 2. Comparison of existing technologies supporting off-network communication.

with a channel bandwidth of 1.4 MHz–20 MHz, and the access method is single-carrier frequency-division multiple access (SC-FDMA). The channel coding scheme of LTE-V2X sidelink is turbo code. The modulation scheme is quadrature phase shift keying (QPSK) or 16-quadrature amplitude modulation (16-QAM), and to support higher transmission rates, 3GPP added 64-QAM in Release 15. The transmit power of LTE-V2X mobile devices is usually set to 23 dBm, supporting a transmission rate of 30 Mb/s, and the transmission delay is generally in the range of 10–100 ms. LTE-V2X Mode 4 uses sensing-based semi-persistent scheduling (SPS) for resource reservation.

NR-V2X [14]: 3GPP proposed NR-V2X in Release 16 to supplement LTE-V2X in scenarios requiring high throughput, ultra-reliability, and low latency, and they will coexist and supplement each other in the future. Similar to LTE-V2X Mode 4, NR-V2X Mode 2 also can directly communicate through sidelink.

The frequency bands of NR include frequency range 1 (FR1) and frequency range 2 (FR2). The bands used by NR-V2X sidelink belong to FR1, which are 5.9 GHz (n47) and 2.5 GHz (n38), respectively. The channel coding scheme of NR-V2X sidelink is low density parity check (LDPC) code. Compared to LTE-V2X, the modulation scheme 256-QAM is added to support higher rate transmission. NR-V2X supports a channel bandwidth of 5-40 MHz, with a maximum transmission rate of 1 Gb/s and an end-to-end latency as low as 3 ms [15]. NR-V2X Mode 2 also uses the sensing-based SPS resource reservation scheme.

Analysis of latency and transmission rate: GSM-R, TETRA, and DMR may have a latency of hundreds of milliseconds or even seconds, while LTE-V2X can achieve end-to-end latency of about 20 ms, and NR-V2X is even more advanced, reaching about 10 ms [15]. As shown in Table 1, most scenarios require end-to-end latency within 100 ms, and some critical scenarios even have a strict requirement of 10 ms. Therefore, GSM-R, TETRA, and DMR may not satisfy the latency requirements of these use cases, while LTE-V2X and NR-V2X may satisfy them most of them.

In addition, GSM-R, TETRA, and DMR are narrowband technologies, and they have low data rates, which are insufficient for future off-network use cases of railways. On the contrary, the data rate performance of LTE-V2X and NR-V2X may meet the data rate requirements, which may reach 30 Mb/s and 1 Gb/s, respectively.

Besides end-to-end latency and data rate, the reliable communication distance of each existing technology is also a key performance indicator that still lacks a comprehensive evaluation. Therefore, the following section provides a detailed simulation analysis of the reliable communication distance.

Reliable Communication Distance Evaluation of Existing Technologies

This section uses Matlab to evaluate the reliable communication distance of each existing technologies in an open field environment. The simulation sets up two trains, one in front of the other, with overhead line masts at regular intervals on the track, in addition to adjacent buildings, trees, and walls. The channel model uses the GSCM consisting of line of sight (LoS) components and multipath components (MPCs). More detailed channel infor-

When the transmit power is increased to 46 dBm, which is usually used for the base station, the existing technologies in the open field environment can satisfy the reliability requirements of 99.99 percent for 3 km communication distance and 99.9999 percent for 1.5 km communication distance.

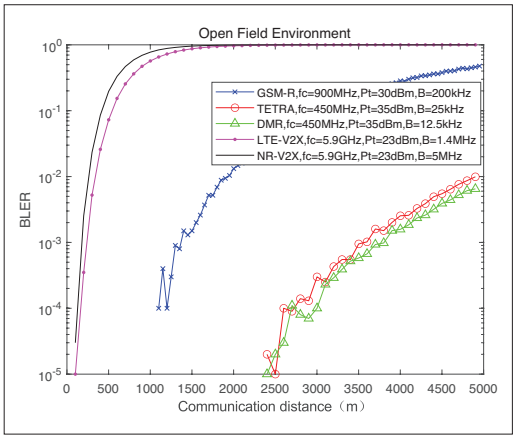


FIGURE 2. BLER performance in an open field environment with a single antenna and without retransmission when the physical layer parameters are specified in the standard.

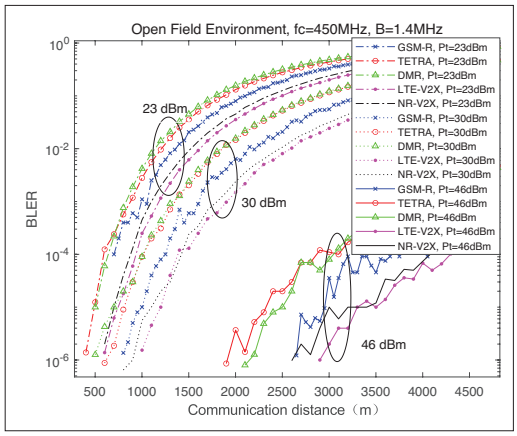


FIGURE 3. BLER performance in an open field environment with a single antenna and without retransmission when f_c = 450 MHz, B = 1.4 MHz.

mation and parameters can be found in [9].

The simulation evaluates the link-level performance of data channels for existing technologies considering the process of random data generation, channel coding, modulation, pass-through channel, demodulation, and decoding. Assume that radio resources are sufficient and both trains can successfully reserve resources. The packet size is set to 256 octets [4], and each technology uses the modulation and channel coding schemes summarized in Table 2, noting that LTE-V2X and NR-V2X use QPSK. In addition, the code rates of the channel coding schemes for all technologies are set to 1/3 for fairness.

Reliable Communication Distance of Different Technologies With Standard Parameters

Figure 2 shows the transport block error rate (BLER) performance of each technique at different communication distances. Each technique uses the frequency, transmission power, and channel bandwidth given by the standard and shown in the legend. Combining Table 1 and Fig. 2, we can get the following conclusions:

Figure 2 shows that LTE-V2X and NR-V2X have a reliable communication range of about 100 m. In contrast, GSM-R has a reliable transmission range of about 1 km, and TETRA and DMR have a reliable communication range of over 2.5 km. The reason is that LTE-V2X and NR-V2X have a low standard power of only 23 dBm and operate at 5.9 GHz. Higher frequency leads to more attenuation and therefore cannot satisfy the reliability and communication distance requirements of the use cases in Table 1. Thus, suitable transmission power and frequency are necessary.

According to the performance analysis earlier and the results in Fig. 2, the existing technologies with the parameters given by the standard cannot fully satisfy the requirements of the use cases in Table 1. In conclusion, to satisfy the requirements of off-network use cases in FRMCS, new technology or evolution of the existing technology is needed.

Impact of Different Parameters

Based on the results of Fig. 2, it is better to choose a lower frequency and an appropriate transmission power to achieve reliable transmission over long distances. Figure 3 simulates the BLER performance of existing technologies at different communication distances by setting the frequency to 450 MHz, the bandwidth to 1.4 MHz, and the transmit power to 23 dBm, 30 dBm, and 46 dBm, respectively.

Figure 2 shows that the BLER performance of LTE-V2X and NR-V2X with 5.9 GHz is inferior to other low-frequency technologies. While Fig. 3 sets the frequency of LTE-V2X and NR-V2X to 450 MHz, their performance is significantly improved, even better than GSM-R, TETRA, and DMR. The reason might be that the LDPC code used in NR-V2X and the turbo code used in LTE-V2X have more significant coding gain than the convolutional, RCPC, and trellis codes used in GSM-R, TETRA and DMR, respectively, but also might introduce additional latency through coding delays, which will be further investigated in future work. Therefore, it is suggested to operate LTE-V2X or NR-V2X at a lower frequency to achieve long-distance transmission.

As shown in Fig. 3, with the same settings of other parameters and the higher transmission power, better performance is obtained. When the transmit power is increased to 46 dBm, which is usually used for the base station, the existing technologies in the open field environment can satisfy the reliability requirements of 99.99 percent for 3 km communication distance and 99.9999 percent for 1.5 km communication distance as seen in Table 1. But the open field environment is only one realistic scenarios, and there are also non-line-of-sight (NLoS) environments where greater fading is caused by obstruction. The performance of these techniques under NLoS environments would be worse than that in an open field environment. Therefore, even if the power is increased to 46 dBm, they may not satisfy the requirements of NLoS scenarios. Hence, blindly increasing the transmission power is not feasible. Instead, if the existing LTE-V2X or NR-V2X is further enhanced, 30 dBm may be a more suitable power, which is now commonly used for train communication.

According to the analysis in Fig. 3, when using the 30 dBm power commonly used in trains, LTE-V2X and NR-V2X are insufficient to support the use cases in Table 1 in terms of reliable communication distance even with lower frequency. However, appropriate enhancement of LTE-V2X

and NR-V2X may have the potential to fulfill these strict requirements. Therefore, besides adjusting the frequency and transmission power, we will further enhance them by adding multiple-input multiple-output (MIMO) and a retransmission scheme in the next part and simulate them at the possible frequency bands of FRMCS to analyze their performance.

Reliable Communication Distance of LTE-V2X and NR-V2X at FRMCS Frequency Band with MIMO and Retransmission

Figure 4 simulates the LTE-V2X and NR-V2X technologies at the possible future FRMCS bands 900 MHz, 1.9 GHz, and 2.1 GHz, and adds 450 MHz for comparison with Fig. 3. 2 × 2 MIMO is used to improve the performance further assuming that the antenna spacing is large enough so that antennas are not correlated with each other. The minimum mean square error (MMSE)-based beamforming technique is used, and the 2 × 2 channel coefficients are generated by the extension of GSCM. In addition, since LTE-V2X only supports blind retransmission, Fig. 4 uses three blind retransmissions, and the transmission power is set to 30 dBm.

In Fig. 3, when the transmission power is 30 dBm with the frequency of 450 MHz, LTE-V2X and NR-V2X can only achieve a reliability performance of 99.9999 percent at less than 1 km communication distance. With the addition of MIMO and retransmission, Fig. 4 shows that they can achieve a reliability performance of 99.9999 percent over 6 km communication distance at 450 MHz in the open field environment, which is a significant improvement and is sufficient for most of the off-network communication use cases in Table 1.

In addition, by adding MIMO and a retransmission mechanism, the performance of LTE-V2X at 450 MHz, 900 MHz, 1.9 GHz, and 2.1 GHz can satisfy the reliability requirements of 99.99 percent for 3 km communication distance and 99.9999 percent for 1.5 km communication distance, as mentioned in Table 1. NR-V2X also can satisfy these requirements at 450 MHz and 900 MHz. However, all performance evaluations are based on an open field environment. For NLoS scenarios, further evaluations will be done in future work.

In conclusion, enhancement of LTE-V2X and NR-V2X by adding MIMO and retransmission has the potential to satisfy all the off-network use cases' requirements in Table 1 at 900 MHz and 450 MHz except for the trackside maintenance warning system. For the consideration of policy permission, 900 MHz might be a more suitable frequency band for FRMCS off-network communication. The 900 MHz currently used by GSM-R may be refarmed for off-network communication in FRMCS.

Opportunities and Challenges

In this section, some potential research directions and challenges are presented to provide insight for future research.

System Parameters Consideration: Frequency, transmission power, bandwidth, and modulation coding scheme significantly affect the performance of the system. A reasonable setting is required to meet the system requirements.

Multi-Air Interface and Multi-Antenna: When the off-network communication mode is introduced, the placement and orientation of the antenna cannot be well optimized for train-to-ground communication, and at the same time for train-to-train communication. When performing mode switching, antenna adjustment will inevitably cause communication interruptions. Similar to the Uu interface (on-network mode) and PC5 (sidelink) interface (off-network mode) in LTE-V2X/NR-V2X, adding a new air interface for railway off-network mode is a good solution (the two air interfaces are independent of each other and can operate simultaneously). The existing Uu-PC5 seamless switching optimization scheme has been relatively mature, which has great reference value for future research on mode switching in railway.

Multiple antennas need a trade-off between diversity and multiplexing, and the high-speed movement of trains can make channel estimation inaccurate. Another key challenge is how to optimize and deploy the antennas.

MAC Protocol Design: If V2X is used for FRMCS off-network communication, the higher transmit power of the PC5 interface will make coexistence with cellular link (Uu) in the same band challenging. In addition, since there is no centralized control center for off-network communication, suppressing interference from other users in the same frequency band and avoiding collisions when reserving the channel are also key challenges. The power control algorithm, channel division protocol, and random access protocol of the MAC layer can be used to avoid these problems. Therefore, the MAC protocols dedicated to the railway environment need to be designed for FRMCS.

Channel Measurement: Accurate channel measurement can help better understand the signal propagation conditions, which can benefit system design and parameter setting. There have already been works on T2T channel measurement, such as [7–9]. However, the future railway off-network communication scenarios are diverse, and there might be several different potential frequencies. Hence, more channel measurements can be done for different scenarios with different frequencies.

Off-Network Communication Assisted by Relay or Satellite Communications: Some extreme scenarios, such as the trackside maintenance warning system, which requires reliability of 99.9999 percent within a communication range of 8.5 km, need some other technologies

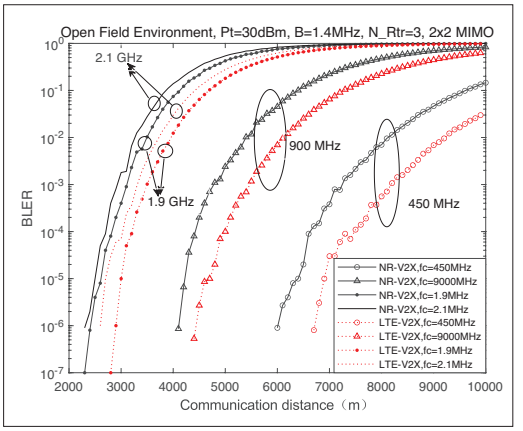


FIGURE 4. BLER performance in an open field environment with MIMO and retransmission when $B = 1.4$ MHz, $P_t = 30$.

An accurate channel measurement can help better understand the signal propagation conditions, which can benefit system design and parameter setting.

According to the performance analysis of different technologies, it is challenging for the existing technologies to fully satisfy the requirements of these cases. However, with the help of MIMO and retransmission scheme, we point out that enhancement on LTE-V2X or NR-V2X operating at 900 MHz has the potential to satisfy the requirements of off-network use cases in FRMCS.

to assist communication. Existing technologies cannot achieve the long-range reliable communication requirements of a trackside maintenance warning system with only one-hop transmission. Relay technology for multihop communication can be adopted to improve communication distance and reliability. In addition, satellite communication technology can also be used to support this use case.

Mission-Critical Framework (MCX): The FRMCS system builds on the 3GPP MCX framework, which complements the transport technology by functions for authentication, functional addressing, and so on. A key challenge is that the MCX framework is typically centralized and has not been used for off-network communications as of today. 3GPP expects to use ProSe direct communication to provide MCX, but lacks a set of standards and requires a lot of future research.

System-Level Simulation: This article performs link-level simulations of the data channel for existing technologies. In the future, more comprehensive system-level simulations are needed to evaluate the overhead and latency, including evaluation of the MAC layer as well as control channel overhead, among others.

Coexistence of Multiple Technologies: Most of the existing railway communications are based on GSM-R, and it will be a long process to complete the transition from GSM-R to its successor (e.g., LTE-R or 5G-R), so they will coexist for a long time in the future. The coexistence of multiple technologies is also a challenge in terms of interference and allocation of bandwidth resources.

Conclusion

This article briefly introduces off-network use cases in FRMCS and summarizes the performance requirements of these use cases. Then we provide an overview of existing technologies (GSM-R, TETRA, DMR, LTE-V2X, and NR-V2X) that may support off-network communication and discuss their physical layer characteristics. In addition, we simulate and evaluate the data channel of existing technologies and compare their performance with the requirements of off-network use cases in FRMCS. According to the performance analysis of different technologies, it is challenging for the existing technologies to fully satisfy the requirements of these cases. However, with the help of MIMO and a retransmission scheme, we point out that enhancement of LTE-V2X or NR-V2X operating at 900 MHz has the potential to satisfy the requirements of off-network use cases in FRMCS. Finally, some future research directions are suggested, including system parameters consideration, multi-air interface and multi-antenna, channel measurement, off-network communication assisted by relay or satellite communications, a mission-critical framework, MAC protocol design, system-level simulation, and coexistence of multiple technologies.

Acknowledgments

This work is jointly supported by the NSFC project (grants No. 61971359 and No.62271419), the Chongqing Municipal Key Laboratory of Institutions of Higher Education (grant No. cquptmct-202104), the Fundamental Research Funds for the Central Universities, Sichuan Science and Technology Project (grant no. 2021YFQ0053), and the State Key Laboratory of Rail Transit Engineering Informatization (FSDI).

References

[1] UIC FU-7100 v. 5.0.0, "Future Railway Mobile Communication System; User Requirements Specification," Feb. 2020.
[2] UIC MG-7900 v. 2.0.0 "Future Railway Mobile Communication System; Use Cases," Feb. 2020.
[3] 3GPP TR 22.889 v. 17.4.0 Rel. 17, "Study on Future Railway Mobile Communication System," Mar. 2021.
[4] 3GPP TR 22.990 v1.0.0 Rel. 18 "Study on Off-Network for Rail," Sept. 2021.
[5] J. Zhao *et al.*, "Future 5G-Oriented System for Urban Rail Transit: Opportunities and Challenges," *China Commun.*, vol. 18, no. 2, Feb. 2021, pp. 1–12.
[6 X. Wang *et al.*, "Train-Centric CBTC Meets Age of Information in Train-to-Train Communications," *IEEE Trans. Intelligent Transportation Systems*, vol. 21, no. 10, Oct. 2020, pp. 4072–85.
[7] A. Lehner, T. Strang, and P. Unterhuber, "Train-to-Train Propagation at 450 MHz," *2017 11th Euro. Conf. Antennas and Propagation*, 2017, pp. 2875–79.
[8] P. Unterhuber *et al.*, "Wide Band Propagation in Train-to-Train Scenarios — Measurement Campaign and First Results," *2017 11th Euro. Conf. Antennas and Propagation*, 2017, pp. 3356–60.
[9] P. Unterhuber *et al.*, "Geometry-Based Stochastic Channel Model for Train-to-Train Communication in Open Field Environment," *2022 16th Euro. Conf. Antennas and Propagation*, 2022, pp. 1–5.
[10] ETSI TS 100 573 v. 6.1.1 "Digital Cellular Telecommunications System (Phase2+) (GSM); Physical Layer on the Radio Path; General Description," July 1998.
[11] ETSI EN 300 396-2 v. 1.4.1 "Terrestrial Trunked Radio (TETRA); Technical Requirements for Direct Mode Operation (DMO); Part 2: Radio Aspects," Dec. 2011.
[12] ETSI TS 102 361-1 v. 2.5.1 "Digital Mobile Radio (DMR) Systems; Part 1: DMR Air Interface (AI) Protocol," Oct. 2017.
[13] 3GPP TS 36.211 v. 14.8.0 Rel. 14 "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation," Sept. 2018.
[14] 3GPP TS 38.211 v. 16.7.0 Rel. 16 "NR; Physical Channels and Modulation," Sept. 2021.
[15] S. Chen *et al.*, "A Vision of C-V2X: Technologies, Field Testing, and Challenges with Chinese Development," *IEEE IoT J.*, vol. 7, no. 5, May. 2020, pp. 3872–81.

Biographies

Jiewen Hu is currently pursuing a Ph.D. degree in the School of Information Science and Technology, Southwest Jiaotong University, China.

Gang Liu [M'15] (gangliu@swjtu.edu.cn) is currently an associate professor in the School of Information Science and Technology, Southwest Jiaotong University.

Yongbo Li is currently pursuing a Ph.D. degree in the School of Information Science and Technology, Southwest Jiaotong University.

Zheng Ma [M'07] is currently a professor in the School of Information Science and Technology, Southwest Jiaotong University.

Wei Wang is currently a senior engineer at the State Key Laboratory of Rail Transit Engineering Informatization (FSDI), China.

Chengchao Liang [S'15, M'17] is currently a professor with the School of Communication and Information Engineering, CQUPT.

F. Richard Yu [S'00, M'04, SM'08, F'18] is currently a professor at Carleton University, Ottawa, Ontario, Canada.

Pingzhi Fan [M'93, SM'99, F'15] is currently a distinguished professor at Southwest Jiaotong University.

It's Time to RENEW Your Membership!

IEEE ComSoc
IEEE Communications Society
IEEE

Empowering communications technology professionals around the globe!

IEEE Communications Society (ComSoc) members belong to a global community of 29,000+ engineers, practitioners and academics working together to advance communications technology for the betterment of humanity.

As an organization, we share expertise, learn, and collaborate in an effort to solve today's communications technology challenges, and create tomorrow's improved capabilities.

As a ComSoc member, you receive exclusive benefits to help you achieve your professional goals and stand out from your peers.

ComSoc Member Benefits:

Career Resources | Professional Networking | Publishing Opportunities | Access to High-Quality Technical Content | Professional Development and Training Options | Conference Discounts | Volunteer Opportunities | Awards Recognition | and more!

Our community is *your* community. JOIN or RENEW today!

www.comsoc.org/membership.

Toward Industry 5.0: Intelligent Reflecting Surface in Smart Manufacturing

Md. Noor-A-Rahim, Fadhil Firyaguna, Jobish John, M. Omar Khyam, Dirk Pesch, Eddie Armstrong, Holger Claussen, and H. Vincent Poor

The authors provide an overview of IRS technology and then conceptualize the potential for IRS implementation in a future smart manufacturing environment to support the emergence of Industry 5.0 with a series of applications.

Abstract

Industry 5.0 envisions close cooperation between humans and machines requiring ultra-reliable low-latency communications (URLLC). The intelligent reflecting surface (IRS) has the potential to play a crucial role in realizing wireless URLLC for Industry 5.0. IRS is forecast to be a key enabler of 6G wireless communication networks as it can significantly improve wireless network performance by creating a controllable radio environment. In this article, we first provide an overview of IRS technology and then conceptualize the potential for IRS implementation in a future smart manufacturing environment to support the emergence of Industry 5.0 with a series of applications. Finally, to stimulate future research in this area, we discuss the strength, open challenges, and opportunities of IRS technology in modern smart manufacturing.

Introduction

Smart manufacturing aims to increase productivity and efficiency by integrating the physical world with the cyber world through the Industrial Internet of Things (IIoT). IIoT now connects millions of industrial devices embedded in the physical world to the Internet (or an organization's intranet) and allows for the integration of data generated by them into information systems and business processes and services. This framework is part of the broader trend known as Industry 4.0. The integration of physical and cyber worlds as part of Industry 4.0 is turning traditional industrial automation and control systems into cyber-physical manufacturing systems.

A major driver in the transformation from industrial automation and control into cyber-physical manufacturing systems is the introduction of private 5G wireless networks into industrial environments [1]. Looking further ahead, it is anticipated that the shift from 5G to 6G will also stimulate a transition from Industry 4.0 towards Industry 5.0. While Industry 4.0 will see enhanced introduction of robotics, we consider Industry 5.0 as the next level of human/automation collaboration, where humans and machines share the work, safely and seamlessly, rather than machines replacing humans. In spite of the fact that robots are more reliable than humans and can do more work, they lack many of the fine motor skills that humans have as well as adaptability and critical thinking skills. In the Industry 5.0 era, robots/machines will be used for repetitive, monotonous, dirty, heavy-duty tasks that represent health hazards for humans (e.g., repetitive strain injury, mental health issues, physical injury). This will free humans up to engage in more stimulating and interesting work, which is harder to automate and requires critical thinking. Thus, Industry 5.0 requires unprecedented collaboration between increasingly powerful and precise machinery and the unique creativity of human beings [2]. 6G is expected to develop better integration of automatic and high-precision manufacturing processes as well as integrating machines and humans into control loops through low latency and high reliability [3].

Compared to traditional wireless communications, industrial wireless communications are already challenged due to metallic structures, electromagnetic interference (e.g., from electrical motor drives or welding apparatus), arbitrary movement of objects (robots and vehicles), room dimensions, or thick building structures. On the other hand, full industrial automation requires ultra-reliability and low-latency communications in order to deliver sensor data and actuation commands at precise instants with designated reliability (i.e., to perform mission-critical industrial processes). Collaboration of humans and machines in Industry 5.0 will add more complexity to the industrial wireless communication system. In addition, the rising demand for many emerging services in innovative industries including augmented/virtual reality (AR/VR) maintenance, holographic control display systems, and so on will bring forth new communication challenges to industrial networks [3]. To meet the communication requirements of such services, Industry 5.0 needs to support advanced technologies such as millimeter-wave/terahertz (mmWave/THz) communications, advanced localization, and efficient energy harvesting in complex industrial environments. The key communication system requirements[1] for Industry 4.0 and Industry 5.0 are presented in Table 1.

A recently developed concept called intelligent reflecting surfaces (IRSs) can serve as a potential solution to many of the above challenges in future

This work was supported by Science Foundation Ireland under Grant 16/RC/3918 (Confirm Centre for Smart Manufacturing) and Grant 13/RC/2077_P2 (CONNECT: The Centre for Future Networks & Communications).

[1] In deriving these requirements, we reviewed which applications would most likely be used in future industrial environments and what requirements they will pose on the communication system. These include asset tracking, mobile robots, and condition monitoring, which are all already considered for Industry 4.0 [1], while applications like brain controlled machinery, AR/VR maintenance, immersive collaborative robotics, augmented clothing, and remote haptic interactions are considered for Industry 5.0 [2].

Md. Noor-A-Rahim, Fadhil Firyaguna, Jobish John, Dirk Pesch, and Holger Claussen are with University College Cork, Ireland; M. O. Khyam is with the Central Queensland University, Australia; E. Armstrong is with Johnson & Johnson, Ireland; H. V. Poor is with Princeton University, USA.

Digital Object Identifier: 10.1109/MCOM.001.2200016

smart manufacturing. An IRS is technology that can significantly improve wireless network performance by creating a programmable radio propagation environment. An IRS is a programmable meta surface containing a large amount of small, low-cost passive antenna arrays that can control a propagating wave's phase, amplitude, frequency, and even polarization. It can increase the efficiency of the wireless network in terms of data rate, coverage, and connectivity. For instance, if a line-of-sight (LoS) link is blocked, an IRS can create a reflective link (i.e., a virtual LoS) to bypass the obstacles between communicating devices.

This article aims to conceptualize the potential for IRS implementation in a smart manufacturing environment to support the emergence of Industry 5.0. The system model is introduced. Several IRS applications specifically relevant to smart manufacturing are presented. We then outline significant future research directions related to the challenges and opportunities associated with the use of IRSs in modern smart manufacturing, and provide conclusions.

Intelligent Reflecting Surfaces

The concept of an IRS is drawn from the concept of a meta surface, which is a 2D form of meta material. Generally, depending on its structural parameters, this engineered man-made material exhibits unique electromagnetic properties that cannot be obtained with conventional materials. The IRS is constructed with a large reconfigurable array of passive sub-wavelength-scale scattering elements (dielectric or metallic patches) that are printed on a grounded dielectric substrate. The size of these patches and their inter-element spacing is usually half of the wavelength or smaller (5 to 10 times smaller) [4].

A meta material unit or patch element is capable of adjusting the phase and amplitude of a reflected signal. The direction of reflected signals from each of these elements can be adjusted in the desired fashion (so as to interfere constructively or destructively at the intended location) by controlling their reflection coefficients in real time. This phenomenon can be characterized by the concept of reflectivity, which is defined as the ratio of the reflected signals to the incident signals. The reflectivity of the meta material unit can be obtained by its state, the incident angles, and reflected angles. The reconfigurability of the meta material units or patch elements are achieved with the help of tunable low-power electronic circuit elements such as positive-intrinsic-negative (PIN) diodes or varactor diodes or radio frequency (RF) switches as shown in Fig. 1 [4]. By controlling the bias voltages of the PIN diodes, each PIN diode can be tuned between ON and OFF states, which facilitates determining the state of the meta material unit [5]. In order to program or configure a smart surface or patch elements remotely, an IRS is equipped with a controller (Fig. 1). The controller is connected to a base station (BS) or access point (AP) to receive relevant control and reconfiguring commands. Although it is not explicitly shown in Fig. 1, an IRS can also be equipped with sensors to help estimate wireless channel conditions [4].

Although IRS technology is promising for application in smart industries, it is still immature, and there are several open issues that should be addressed including theoretical design and practical integration and engineering manufacture of IRSs. A comprehensive study in this direction can be found in [6]. Also, we discuss some open issues associated with the application of IRSs in modern smart manufacturing.

Key performance indicator	Industry 4.0	Industry 5.0
Data rate	Up to 10 Gb/s	Up to 100 Gb/s
Latency	100 ms to 250 μs	Less than 100μs
Reliability (packet error rate)	10^{-5} to 10^{-8}	Up to 10^{-10}
Connectivity density	1 device/m^2	10 devices/m^2
Energy efficiency in communications	1×	10× that of Industry 4.0

TABLE 1. Key communications requirements for Industry 4.0 and Industry 5.0.

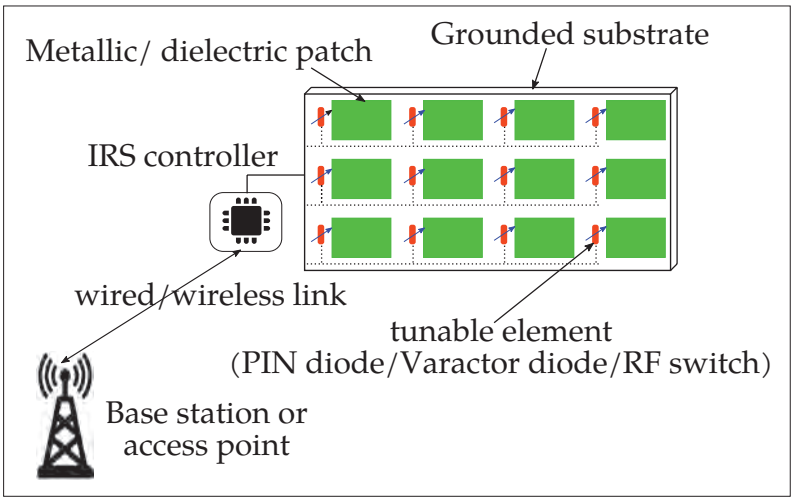


FIGURE 1. IRS architecture.

Applications of IRS in Smart Manufacturing

In this section, we present a number of use cases for an IRS or multiple IRSs in a future manufacturing environment with many autonomous and mobile devices, showing where their functionality can be of particular use.

Blockage Mitigation

Obstructions and blockage are major issues for signal coverage and can cause intermittent and poor connectivity. To circumvent obstructions, an IRS can help steer the incident signals around an obstruction and cover the area shadowed from the base station. With its large number of passive reflective elements, the IRS enables the adaptation of a wireless environment to overcome the blockage and provide a strong reflective non-line-of-sight (NLoS) link.

In the following, we show the path loss characteristics when using an IRS with a transmission frequency of 30 GHz. We assume that the communication link between BS and receiver is completely blocked and an IRS is set up to circumvent the blockage, as depicted in Fig. 2. In our scenario, we assume that the IRS is composed of $N \times N$ elements which are positioned 20 m away from the BS in such a way that the IRS has an LoS link with both BS and receiver. Figure 3 shows the end-to-end path loss as a function of the distance d between the IRS and the receiver. This path loss

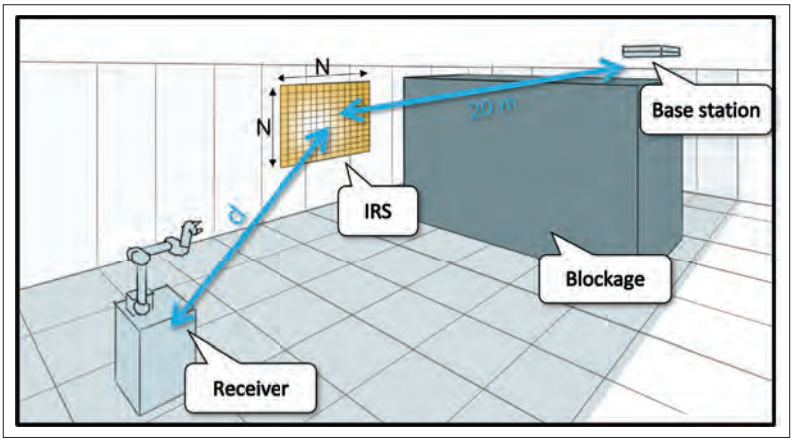


FIGURE 2. Scenario for IRS path loss analysis.

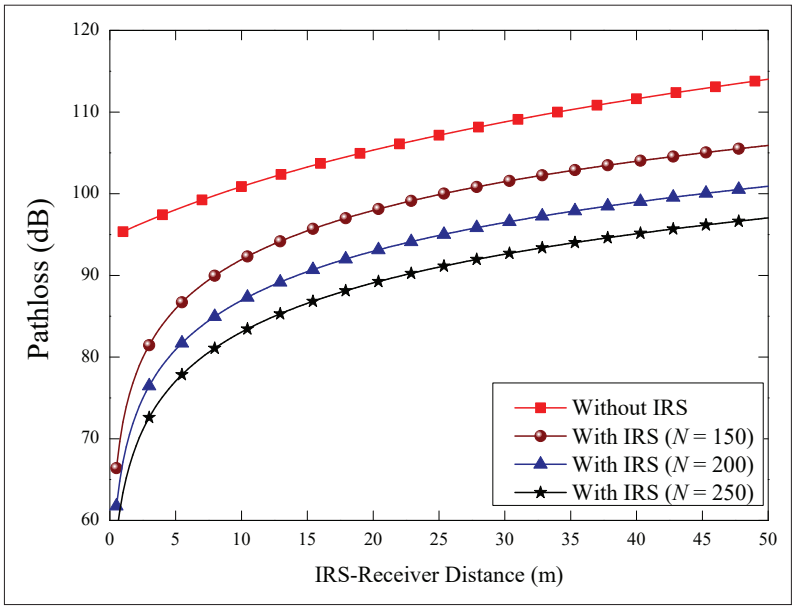


FIGURE 3. Path loss characteristics of an IRS as a function of the distance *d* between the IRS and the receiver.

is calculated using an IRS-based path loss channel model presented in [7]. We have also plotted path loss for a scenario without IRS, which is calculated using the Third Generation Partnership Project (3GPP) indoor factory NLoS channel model with a dense cluttered environment. Significant reduced path loss is observed with IRS scenarios compared to the scenario without IRS. With varying numbers of IRS elements, the figure shows that for any given IRS-receiver distance, over 10 dB received power gain can be achieved by increasing *N* from 150 to 250.

This is a positive result to enable flexible smart manufacturing environments, where a plant reconfiguration can degrade a given area from being covered by wireless signals to one that is shadowed by machinery. Hence, the signal coverage can be automatically adapted by steering reflecting beams as the factory plant is reconfigured without the need to redeploy network infrastructure, as illustrated in Fig. 4. Furthermore, to provide the best radio environment, the IRS could be mounted in a mobile support unit, so its position can be optimized according to the new floor layout. This flexibility provided by IRSs is key to enable features and applications that rely on high coverage availability described in the following, thus improving communication and operational performance of a smart factory.

Millimeter-Wave and Terahertz Communications

Wireless communication for smart manufacturing is characterized by strict link and system requirements regarding availability, reliability, and latency. For instance, a closed-loop motion control use cases may demand cycle times lower than 1 ms and 99.9999 percent service availability for more than 100 nodes [8]. To support such application requirements, mmWave or THz spectrum will provide wireless channels with wide bandwidth to accommodate large numbers of nodes operating with high data rate and low latency. However, propagating mmWave/THz signals suffer from very high path attenuation and lower penetration through materials compared to lower-frequency signals in the currently used sub-6 GHz bands. This means that communication links are more vulnerable to blockage, potentially affecting their performance. Although the use of highly directional antennas can compensate for some of the path loss, narrow beamwidths may make a link even more vulnerable to blockage.

The mmWave/THz signals have wavelengths at the scale of the surface roughness of many objects, which suggests that scattering may not be neglected as it can be at lower frequencies. Also, the scattered power relative to the reflected power at mmWave/THz frequencies increases with the incident angle, and lower reflection losses (e.g., stronger reflections) are observed as frequencies increase for a given incident angle. This means that the signal energy can be more diffused in the environment, and the propagation characteristics are highly dependable on surfaces and incident angles of the incident waves. Here, the application of IRS for mmWave/THz could be a means to generate a more predictable and controllable channel to overcome the effects of scattering. To implement this, we need to jointly optimize transmit beamforming and IRS phase shift parameters, maximizing the received power [9].

Wireless Energy Transfer

The envisioned future smart manufacturing environment will consist of many wireless sensors and actuators traditionally powered by batteries, which deplete over time. Replacing depleted batteries of thousands of sensors and actuators is a costly task. Additionally, many sensors may be installed in sensitive places that are not suitable for frequent battery replacements. The use of wireless energy harvesting, which allows sensors/actuators to harvest power from a signal that was intended for data transmission or power transmission, has been shown to be an excellent solution to address this issue. The problem with wireless power transfer, however, is the drop in wireless power over large transmitter-to-receiver distances. A wide range of techniques have been proposed to overcome this power loss, such as waveform design, energy transmission and scheduling, and energy beamforming, which can be implemented at the transmitter and/or receiver in order to improve the efficiency of wireless power transfer. Unfortunately, many of the above solutions are not ideal for smart manufacturing environments

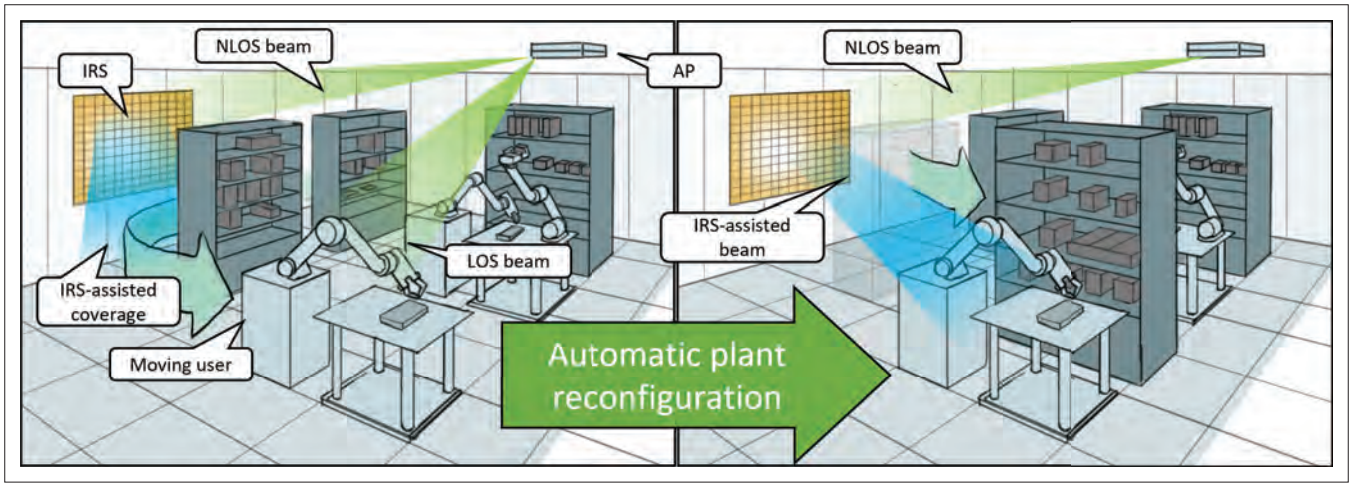


FIGURE 4. Automatic coverage adaptation with an IRS when a factory floor is reconfigured due to production demand.

as they require significant signal processing capabilities, which many low-power IIoT devices do not possess.

Smart manufacturing could benefit from wireless power transfer enhanced by IRSs. Thanks to the deployment of IRSs close to IIoT devices, the issue of high path loss can be effectively alleviated by creating an energy-efficient charging zone for those devices, as depicted on the left of Fig. 5. When IRSs are deployed correctly in LoS with transmitters and receivers and their beamforming capabilities are fully exploited, the received power of nearby IIoT devices can be substantially increased. An energy receiver can take advantage of an IRS's passive beamforming to improve the transmission efficiency of wireless energy while simultaneously enhancing the signal strength at an information receiver. Moreover, it realizes the possibility of improving both the rate and energy performance in wireless power transfer by increasing the wireless charging efficiency. This in turn helps to reduce the transmit power and provides more flexibility in the design of transmit beamforming for information receivers. The effectiveness of passive beamformers for wireless power transfer is expected to be crucial in practice. To achieve their benefits, however, they require channel state information at the energy transmitter.

Sensing and Localization

Sensing and localization in smart manufacturing create the opportunity to continuously monitor individual products, providing the possibility for product customization through tighter control, management, and analysis of critical manufacturing parameters. Thus, acquiring the precise location of objects and being able to sense local information and ambient parameters in the environment in industrial settings are becoming indispensable to enable location- and sensing-based services and applications. For example, the transparent production and logistics processes of smart manufacturing can be improved significantly by knowing what is happening when, where, and how. Automated guided vehicles can improve production supply, assembly lines (through transport platforms), and warehouse logistics systems. IRSs offer significant advantages for precise localization (e.g., using the angle of arrival method) and high-resolution sensing solutions in industrial settings since it can actively customize the propagation environment, as illustrated on the right of Fig. 5. The underlying idea of wireless signal sensing is based on the principle that receivers can identify the effects that sensing targets have on wireless signal propagation. The receiver exploits the observations to detect target behavior. Unlike conventional sensing techniques, IRS-assisted sensing creates a controllable radio environment in preferred directions interacting with sensing targets. As a result, IRS-assisted sensing does not require an LoS link between the receiver and the sensed target [10]. On the other hand, in IRS-assisted localization, an IRS is deployed between the AP and receiver in such a way that the AP can investigate a user's reflected signal through various IRS configurations to achieve accurate locations.

Mobile Edge Computing

IRSs can support mobile edge computing (MEC) in smart manufacturing. The MEC paradigm extends computing resources from the cloud to the network's edge. Future smart factories will have a very large number of wireless devices generating large volumes of data in real time. Generally, these devices do not possess the required processing power and battery capacity, creating the need to offload processing operations to the edge servers (preferred) or cloud platforms. This helps to reduce the end-to-end delay and avoids unwanted network congestion [11].

An IRS can help establish strong wireless computation offloading links, reducing packet losses/retransmissions and enhancing spectral and energy efficiency. As shown in Fig. 5, an IRS-assisted MEC system in a smart factory consists of one or more APs/BSs with co-located edge computing nodes/servers, IRS elements along with its controller, and large numbers of field devices. MEC servers are usually co-located with the APs for ease of joint optimization of computational and communication resources. Usually, in an industrial environment, the APs mounted on the ceiling are connected to the edge servers on the shop floor over a wired connection (e.g., fiber). This limits the flexibility of the smart factory. The application of IRSs, along with upcoming wireless technologies (5G/6G), helps replace this with a completely

In any wireless network, one of the most important tasks is to allocate radio resources optimally. In general, IRS radio resource management is mainly concerned with power allocation, bandwidth allocation, and node-IRS connectivity.

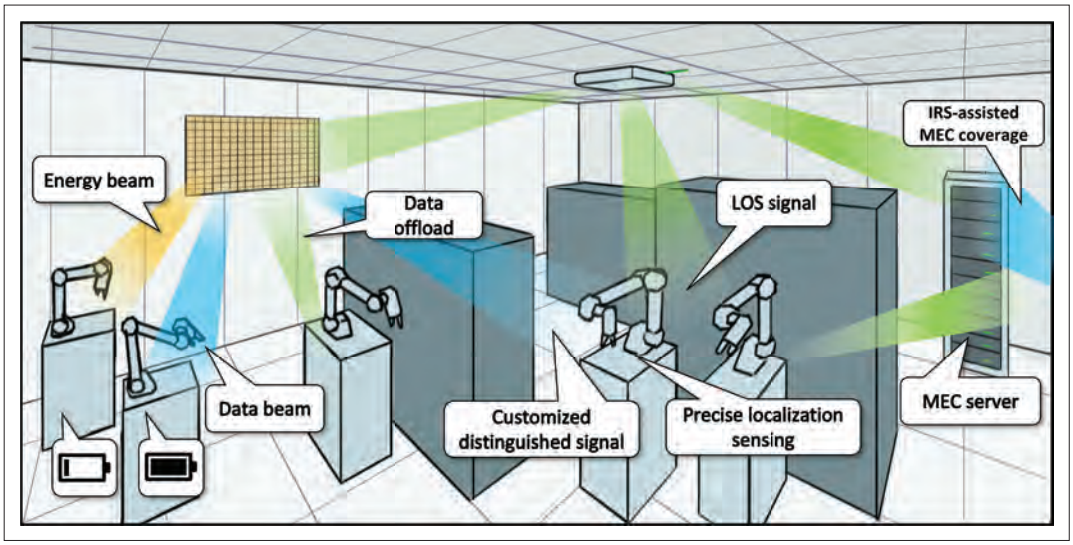


FIGURE 5. IRS-assisted wireless energy transfer, data offloading for mobile edge computing, and high-accuracy localization.

wireless link with ultra-high reliability, low latency, and high bandwidth. There can be several distributed edge computing servers on the shop floor, and the wireless devices can offload their computations to their nearby servers wirelessly. A virtual LoS link with enhanced channel gain can be established between the field devices and the AP/edge server by adequately tuning the IRS reflecting elements, which helps to offload data to the MEC server more quickly. The processed results/control actions are also quickly fed back to end nodes, shortening the end-to-end delay. Currently, in many smart manufacturing applications, local node processing is adopted due to weak communication links, resulting in idle resources at the edge. Thus, IRSs can help to exploit powerful edge computational resources better by making them more accessible to an increased number of wireless field devices. However, carrying out the optimal computational resource allocation at the edge servers along with the communication resource allocations at the BS/AP and the dynamic tuning of IRS reflection coefficients will be a major challenge.

Challenges and Open Issues

Environment-Aware Passive Beamforming

One of the main challenges for the successful application of IRSs in smart manufacturing is designing environment-aware and dynamic passive beamforming. In practice, the design of IRS passive beamforming is determined by the discrete amplitude and phase shift levels of each element. A beam steering process requires coordination of the phase control of individual scattering elements. Despite limited phase shifts available at an individual IRS scattering element, an IRS with a large number of scattering elements can enable more flexible phase tuning. However, computational complexity is a price to pay for such flexibility. Moreover, a larger number of scattering elements means greater difficulty in channel estimation, which could hinder efficient phase control.

While exhaustive search may provide the best solution for determining the best amplitude/phase-shift levels, the approach is computationally complex, and may be infeasible for scenarios where energy savings are paramount. Efficient algorithms are therefore required to estimate channels and control phase shifts of all scattering elements in real time following the dynamics of the radio environment. A practical solution as opposed to exhaustive search can be achieved by solving the problem with continuous amplitude and phase shift values, and then calculating the closest discrete values of the obtained solutions [12].

To realize practical and efficient IRS beamforming, machine learning approaches can assist to effectively resolve the above problems by using locally observed information in the smart manufacturing environment. High numbers of scattering elements and sensors means that a significant amount of information can be collected during channel sensing, facilitating machine learning approaches based on large datasets. The use of data-driven machine learning has the potential to minimize the overhead of information exchange between the IRS and active transceivers. In an IRS-based smart manufacturing environment, however, machine learning approaches must be designed to fit the hardware constraints. Large-scale experimental evaluations are essential to gain more insights into the effectiveness of passive beamforming of IRS in real-world deployments.

Radio Resource Management

In any wireless network, one of the most important tasks is to allocate radio resources optimally. In general, IRS radio resource management is mainly concerned with power allocation, bandwidth allocation, and node-IRS connectivity. Due to the specific dynamics of interference in IRS-enabled wireless environments, transmit power allocation is an essential component for the effective operation of an IRS-enabled environment. In smart manufacturing, numerous wireless devices embedded in machines, autonomous vehicles, and the environment coexist in close proximity, making this an even more pressing problem. In order to minimize interference while maximizing the system's capacity, effective power allocation approaches need to be developed. On the other hand, bandwidth allocation determines the most suitable allocation of users to different sub-chan-

nels to increase bandwidth efficiency. Due to the frequency-agnostic nature of IRS elements, one common IRS reflection matrix is shared among sub-channels, making optimization problematic. In order to address the problem, dynamic passive beamforming can be used. In this scheme, the resource blocks are dynamically assigned to different user groups with different IRS phase shifts for different time slots.

Another major challenge in enhancing the wireless network performance using IRSs lies in associating users/wireless devices to IRSs and selecting their communication mode. Some wireless devices may have an excellent direct LoS link with the BS and hence need not be associated with any IRS. Other devices may make use of single reflection links for better network performance and need to be associated with either the user-side or BS-side IRS, while still others may take advantage of both the single reflection and double reflection links. Most of these wireless devices in a smart factory are highly mobile, which results in a highly challenging dynamic IRS-user association. To assign users optimally to different IRSs, the channel state information (CSI) of all communication links is essential, which is very difficult to obtain in practice. How beneficial it is to integrate sensing devices onto IRSs for channel sensing, making it semi-passive, is another question that needs to be investigated. Another challenge lies in establishing a reliable wireless communication link between the IRS controller and the BS. Industry 5.0 forecasts the replacement of the rigid wired communication links in a smart factory, questioning use of wired backhaul links between IRSs and BSs, especially when they are distributed across the shop floor.

Channel Characterization

There are two major challenges for end-to-end analysis of an IRS system. First, to analyze the performance limits of an IRS link, new channel propagation models are needed to obtain the link budget analysis. Path loss models depend on several parameters including the size of the IRS and the mutual distances between the transmitter/receiver and the IRS [13].

Second, to decode the signal reflected by the IRS, the channel should be properly estimated. In addition to estimating the direct link between transmitter and receiver, two IRS-assisted channels need to be estimated (i.e., the transmitter-IRS and IRS-receiver channels), and they cannot be separately estimated via traditional training-based approaches in general because IRSs are typically passive and cannot perform channel estimation by themselves. As a result, alternative approaches are needed to perform channel estimation, while keeping complexity and overhead of IRS operations as low as possible [4]. The problem becomes even more challenging with large IRS arrays since the time overhead to perform the channel estimation increases linearly with the number of IRS elements [14]. Furthermore, such channel models and estimation approaches should consider the specific industrial environment. In a factory floor environment, the presence of metallic surfaces on machinery furniture and vehicles leads to a wide range of values for channel parameters, such as path loss and multipath parameters, especially when using high frequency signals. The applicability of the current/new models (both electromagnetic material models and the IRS-assisted wireless channel models) needs to be validated for the manufacturing environment. Future research should also concentrate on efficient environment-aware dynamic channel characterization approaches.

Deployment Issues

IRSs can be deployed in an industrial environment using different strategies:

- Close to the distributed wireless devices (known as user-side IRS deployment)
- Close to the base station (known as BS-side IRS deployment)
- In a hybrid style combining both user-side and BS-side IRS deployment [15]

Each has its pros and cons. User-side IRS deployment provides enhanced network coverage mainly for the users or wireless devices within its local vicinity. In contrast, the BS-side deployment provides extended network coverage. One of the main motivations for using an IRS is to provide a virtual LoS link between base station and wireless devices whenever obstacles are present. The placement of a user-side IRS is a relatively easy way to establish a virtual LoS link between BS and intended local users, whereas placing a BS-side IRS to establish a virtual LoS link for all its users is difficult. The communication signaling overhead between the IRS controller and the BS for tuning the reflection coefficients of the IRS is relatively low for BS-side deployment due to their close proximity. Hybrid IRS deployment combines the advantages of both user-side and BS-side schemes. It also helps to exploit double reflection links (inter-IRS reflection links) to provide more LoS paths between the served users/wireless devices in a smart factory and the BS/AP. At the same time, a hybrid deployment scheme brings additional complexity in the design, deployment, and management of an IRS. However, the main challenge with these three options is exactly where and how to deploy them. IRSs may be deployed in a centralized or distributed fashion (for a given number of reflecting elements), and it is not yet clear which approach is the best for an industrial environment. An IRS's low cost provides the flexibility to opt for a dense deployment on a factory floor if required. However, their joint network performance optimization will be a challenging task.

Conclusion

In this article, we have discussed the prospects of IRS-aided wireless networks in a smart manufacturing environment to support the evolution toward Industry 5.0 by unfolding their potential features and advantages through different wireless network scenarios. As IRS technology is still in its infancy, we have elaborated on the most pressing challenges as well as the potential opportunities for research into future IRS-aided wireless factory automation. Thus, it is hoped that this article will serve as a useful and inspiring resource for future research on IRS-based smart manufacturing to unlock its full potential in a future industrial environment.

As IRS technology is still in its infancy, we have elaborated on the most pressing challenges as well as the potential opportunities for research into future IRS-aided wireless factory automation.

References

[1] B. Chen *et al.*, "Smart Factory of Industry 4.0: Key Technologies, Application Case, and Challenges," *IEEE Access*, vol. 6, 2018, pp. 6505–19.
[2] P. K. R. Maddikunta *et al.*, "Industry 5.0: A Survey on

Enabling Technologies and Potential Applications," *J. Industrial Info. Integration*, vol. 26, Mar. 2022, p. 100,257.
[3] K. B. Letaief *et al.*, "The Roadmap to 6G: AI Empowered Wireless Networks," *IEEE Commun. Mag.*, vol. 57, no. 8, Aug. 2019, pp. 84–90.
[4] M. Di Renzo *et al.*, "Smart Radio Environments Empowered by Reconfigurable Intelligent Surfaces: How It Works, State of Research, and the Road Ahead," *IEEE JSAC*, vol. 38, no. 11, 2020, pp. 2450–2525.
[5] Q. Wu *et al.*, "Intelligent Reflecting Surface Aided Wireless Communications: A Tutorial," *IEEE Trans. Commun.*, 2021.
[6] E. Basar *et al.*, "Wireless Communications through Reconfigurable Intelligent Surfaces," *IEEE Access*, vol. 7, 2019, pp. 116,753–73.
[7] E. Björnson, Ö. Özdogan, and E. G. Larsson, "Reconfigurable Intelligent Surfaces: Three Myths and Two Critical Questions," *IEEE Commun. Mag.*, vol. 58, no. 12, Dec. 2020, pp. 90–96.
[8] 3GPP, "5G; Service Requirements for Cyber-Physical Control Applications in Vertical Domains," Tech. Rep. TS 122 104 v. 16.5.0, Sept. 2020.
[9] P. Wang *et al.*, "Intelligent Reflecting Surface-Assisted Millimeter Wave Communications: Joint Active and Passive Precoding Design," *IEEE Trans. Vehic. Tech.*, vol. 69, no. 12, 2020, pp. 14,960–73.
[10] J. Hu *et al.*, "Reconfigurable Intelligent Surface Based RF Sensing: Design, Optimization, and Implementation," *IEEE JSAC*, vol. 38, no. 11, 2020, pp. 2700–16
[11] Z. Chu *et al.*, "Intelligent Reflecting Surface Assisted Mobile Edge Computing for Internet of Things," *IEEE Wireless Commun. Letters*, vol. 10, no. 3, 2021, pp. 619–23.
[12] Q. Wu and R. Zhang, "Toward Smart and Reconfigurable Environment: Intelligent Reflecting Surface Aided Wireless Network," *IEEE Commun. Mag.*, vol. 58, no. 1, 2020, pp. 106–12.
[13] W. Tang *et al.*, "Wireless Communications with Reconfigurable Intelligent Surface: Path Loss Modeling and Experimental Measurement," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, 2020, pp. 421–39.
[14] Z. Wang, L. Liu, and S. Cui, "Channel Estimation for Intelligent Reflecting Surface Assisted Multiuser Communications: Framework, Algorithms, and Analysis," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, 2020, pp. 6607–20.
[15] W. M. R. Z. Changsheng You and Beixiong Zheng, "How to Deploy Intelligent Reflecting Surfaces in Wireless Network: BS-Side,User-Side, or Both Sides?," *J. Commun. and Info. Networks*, vol. 7, no. 1, 2022, p. 1.

Biographies

Md. Noor-A-Rahim (m.rahim@cs.ucc.ie) received his Ph.D. degree from the Institute for Telecommunications Research, University of South Australia in 2015. He is currently a research fellow with the School of Computer Science and Information Technology, University College Cork, Ireland.

Fadhil Firyaguna (ff28@cs.ucc.ie) received his Ph.D. degree from Trinity College Dublin, Ireland in 2020. He is a post-doctoral researcher with the School of Computer Science and Information Technology, University College Cork.

Jobish John (j.john@cs.ucc.ie) received his Ph.D. in electrical engineering from the Indian Institute of Technology, Bombay in 2020. He is currently working as a researcher with the School of Computer Science and Information Technology, University College Cork.

Mohammad Omar Khyam (m.khyam@cqu.edu.au) received his Ph.D. degree from the University of New South Wales, Australia, in 2015. He is currently with Central Queensland University, Melbourne, Australia.

Dirk Pesch (d.pesch@cs.ucc.ie) is a professor in the School of Computer Science and Information Technology at University College Cork. He holds a Dipl.Ing. degree from RWTH Aachen University, Germany, and a Ph.D. from the University of Strathclyde, Glasgow, Scotland.

Eddie Armstrong (earmstr1@its.jnj.com) is an engineering fellow with Johnson & Johnson. He received his Ph.D. degree from the Computer Science and Information Systems Department, University of Limerick, Ireland.

Holger Claussen (h.claussen@cs.ucc.ie) is head of the Wireless Communications Laboratory at Tyndall National Institute and a research professor in the School of Computer Science and Information Technology at University College Cork.

H. Vincent Poor (poor@princeton.edu) is the Michael Henry Strater University Professor of Electrical Engineering with the Faculty at Princeton University. He received his Ph.D. degree in electrical engineering and computer science from Princeton University, New Jersey, in 1977.

CALL FOR PAPERS

IEEE COMMUNICATIONS MAGAZINE

FEATURE TOPIC: NETWORKS FOR BLOCKCHAIN-ENABLED APPLICATIONS

BACKGROUND

With fast evolution of computer and networking technologies, more and more systems and applications are based on interconnecting different devices and systems to form the Internet-of-Something: things, vehicles, and even bodies. Many among those systems require permanent, distributed data storage facilities which are being increasingly implemented using replicated, tamper-proof blockchain ledgers. Blockchain technology brings many challenges that include efficient and reliable data distribution, and efficient and collusion-proof consensus mechanisms. Coverage of wide geographical areas brings another set of problems that relate to long propagation delays, flexible dynamic routing, and optimal use of SDN technologies. In addition, a plethora of existing security- and privacy-related constraints on the networks are exacerbated, rather than solved, by blockchain technology.

The purpose of this Feature Topic (FT) is to bring together the latest research results related to analysis, design, and implementation of network and communication systems that support blockchain-enabled Internet-of-Anything systems and applications. To this end, we are seeking paper in one or more of the following areas:

Topics of interest include, but are not limited to:

- Network performance of blockchain applications: data propagation and distribution protocols
- Blockchain and consensus protocols: optimizing the consensus in a networked application
- Protecting the contents of blockchain transactions and blocks
- Authenticating users and nodes in a blockchain-based system
- User and data privacy in blockchain networks
- Encryption techniques for use in blockchain networks and applications
- Challenges of blockchain based wide area networks
- Adapting SDN technologies to blockchain networks
- Blockchain applications in a wireless communication environment
- Internet of Vehicles and blockchain
- Optimizing blockchain operation and storage in resource constrained environments
- Blockchain for Internet of Things systems and applications
- Mobile edge computing and blockchain

SUBMISSION GUIDELINES

Manuscripts should conform to the standard format as indicated in the Manuscript Submission Guidelines in the *IEEE Communications Magazine* website. Please, check these guidelines carefully before submitting since submissions not complying with them will be administratively rejected without review.

All manuscripts to be considered for publication must be submitted by the deadline through Manuscript Central. Select the "FT-2224/Networks for Blockchain-Enabled Applications" topic from the drop-down menu of topics. Please observe the dates specified here below noting that there will be no extension of the submission deadline.

IMPORTANT DATES

Manuscript Submission Deadline: 30 November 2022
First Cycle of Reviews: 28 February 2023
Second Cycle of Reviews and Final Decisions: 30 April 2023
Publication Date: Second/Third Quarter 2023

GUEST EDITORS

Jelena Misic
Toronto Metropolitan University (formerly Ryerson University), Canada
jmisic@ryerson.ca

Vojislav Misic
Toronto Metropolitan University (formerly Ryerson University), Canada
vmisic@scs.ryerson.ca

Haisheng Guo
Chief Engineer at Data Intelligence Platform Department, ZTE Corporation, China
Haisheng Guo

ACCEPTED FROM OPEN CALL

HAPS-ITS: Enabling Future ITS Services in Trans-Continental Highways

Wael Jaafar and Halim Yanikomeroglu

The authors discuss how HAPS systems can enable advanced ITS services for trans-continental highways, presenting the main requirements of HAPS-ITS and a detailed case study of the Trans-Sahara highway.

Abstract

With the advent of rapid globalization and the inter-border supply chain network, the reliability and efficiency of transportation systems have become even more critical. Indeed, trans-continental highways need particular attention due to their important role in sustaining globalization. In this context, intelligent transportation systems (ITS) can actively enhance the safety, mobility, productivity, and comfort of trans-continental highways. However, ITS efficiency depends greatly on the roads where they are deployed, on the availability of power and connectivity, and on the integration of future connected and autonomous vehicles. To this end, high altitude platform station (HAPS) systems, due to their mobility, sustainability, payload capacity, and communication/caching/computing capabilities, are seen as a key enabler of future ITS services for trans-continental highways; this paradigm is referred to as HAPS-ITS. The latter is envisioned as an active component of ITS systems to support a plethora of transportation applications, such as traffic monitoring, accident reporting, and platooning. This article discusses how HAPS systems can enable advanced ITS services for trans-continental highways, presenting the main requirements of HAPS-ITS and a detailed case study of the Trans-Sahara highway.

Introduction

There was a time when traveling within and across countries meant riding for months in horse-drawn wagons or camel-led caravans. With the development of industrialized transportation systems, all of this has changed. Given the widespread availability of automobiles and train travel in the 20th century, a whole new era began, involving overland roads, canals, bridges, and railways. These routes changed our way of life, allowing people and goods to move between cities and across areas previously considered uninhabitable. In the process, the economical, social, cultural, and territorial aspects of globalization has accelerated. One of its major vectors is the development of trans-continental highways. The latter are primary arteries that connect cities within a single country, such as the Trans-Canada highway (7,821 km), Trans-Siberian highway (11,000 km), and Australian Highway 1 (14,500 km), or many cities across different countries, such as the Trans-African highway network (56,683 km).

Another factor driving the development of highway infrastructure is the global market of cellular vehicle-to-everything for intelligent transportation systems (ITS), which is expected to exceed US$1 trillion circa 2030. This will be driven by the growing integration of advanced technologies to manage vehicular traffic in densely populated mega-cites and road infrastructures creaking under the strain of rapid urbanization and growing numbers of vehicles. Also, since road transportation accounts for 27 percent of global CO_2 emissions, the use of ITS tools is advocated to reduce transportation carbon footprint [1]. Indeed, ITS incorporates technologies such as the Internet of Things (IoT), radar, data processing, dissemination/transmission, and intelligent control to improve safety, mobility, productivity, and comfort.

Unlike urban areas, where ITS focuses on productivity, on highways the primary focus is on safety and reducing mortality rates [2]. The issue of fatalities on trans-continental highways is accentuated by the fact that several road segments may be totally isolated. Although road infrastructures have been benefiting from ITS for decades — for example, path finder, dedicated short-range communications (DSRC), and variable message signs (VMSs), a new generation of ITS technologies, such as connected and autonomous vehicles (CAVs), are creating novel applications that demand smart infrastructure involvement.

The lack of connectivity along trans-continental highways is a major limiting factor for ITS goals. Indeed, vehicles rely only on their own intelligence and occasional vehicle-to-vehicle (V2V) communications to improve road safety, establish platooning, and reduce their carbon footprint, while services that require remote control or cloud services cannot be realized for lack of efficient vehicle-to-infrastructure (V2I) communications.

Although satellites can be leveraged for ITS services in any area, their use is limited to delay-tolerant services, while critical ITS applications cannot be handled. To tackle this issue, the use of a constellation of high altitude platform stations (HAPSs) is advocated. In this article, we envision HAPS as the main enabler of future ITS services in trans-continental highways. Indeed, a HAPS node is a wireless network node that operates at a typical altitude of 20 km. Due to recent innovations in autonomous avionics, array antennas, battery, and solar energy, HAPS systems are emerging as a principal component of next-generation networks [3]. In industry, several HAPS startups are leading the way toward high-speed

Digital Object Identifier:
10.1109/MCOM.001.2200025

The authors are with Carleton University, Canada.

0163-6804/22/$25.00 © 2022 IEEE

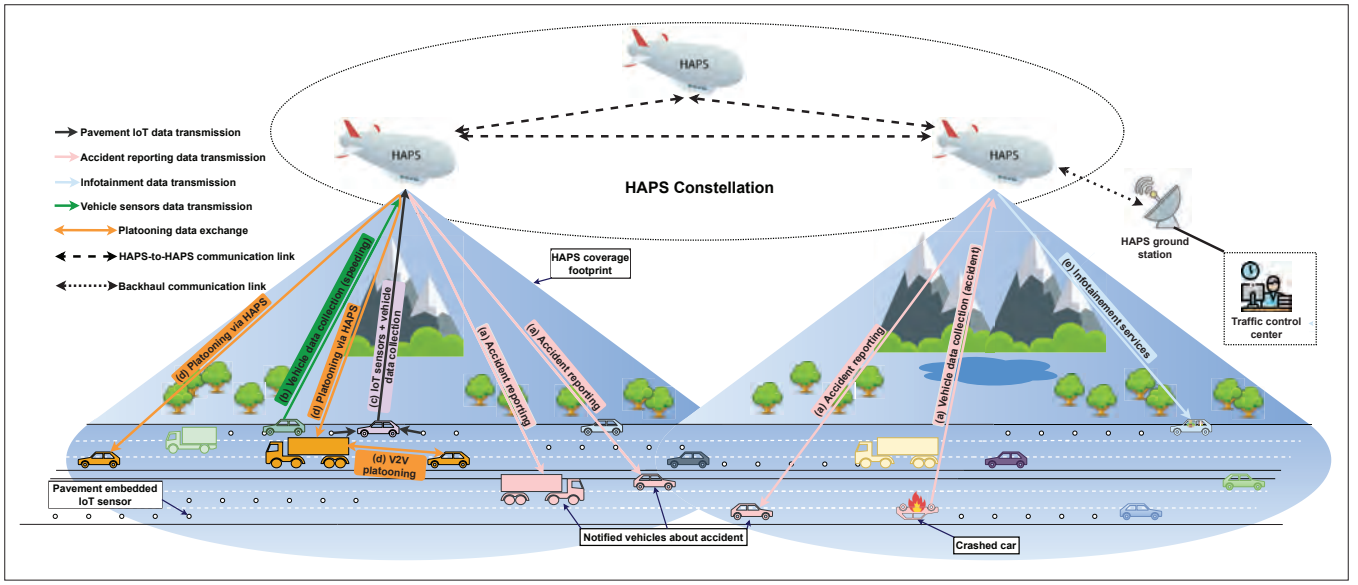


FIGURE 1. Example of data flows for HAPS-supported ITS services.

connectivity from the stratosphere, including HAPSMobile, Thales Alenia Space, and Stratospheric Platforms Limited. As a super macro base station (BS) [4], a HAPS node is expected to provide wireless and Internet connectivity in a wide area up to 500 km in radius (ITU-R F1500), thus enabling several ITS applications like traffic monitoring, accident reporting, platooning, and sensor data collection. Moreover, we propose HAPS as an aerial data center capable of processing big data for ITS services such as road traffic accident analysis, traffic prediction and route planning, and fleet management and control [5]. Similarly, supported by a high storage payload, HAPS caching can be leveraged to provide infotainment services to passengers (e.g., video streaming and gaming). Since communication, computing, and caching resources may not be sufficient in a single HAPS, we envision the use of multiple HAPS nodes with high HAPS-to-HAPS (H2H) data rate links to handle ITS services and improve reliability in failures.

HAPS-ITS for Enhanced ITS Services in Trans-Continental Highways

Given limited government budgets, innovation is needed to get the best from existing infrastructure assets. Also, with the proliferation of CAVs, transportation networks need to be ready with adequate ITS technologies to support CAV services.

ITS Challenges in Trans-Continental Highways

Although new ITS functionalities are attractive for making highways safer, more efficient, and eco-friendly, a number of challenges need to be addressed. First, trans-continental highways suffer from spotty cellular coverage due to harsh terrain conditions (e.g., mountain areas) and from a lack of economic interest from service providers. Consequently, critical ITS services such as accident reporting are inoperable in uncovered areas. Also, real-time infrastructure condition and traffic monitoring cannot be realized in the absence of V2I communications. Since CAV functions are mainly executed by an onboard computer, the latter may potentially fail. In such a situation, the vehicle becomes either "blind" and relies fully on the driver, or leverages V2V and V2I links to offload its tasks to another vehicle or a roadside unit (RSU). Nevertheless, this alternative requires the presence of nearby vehicles and/or cellular coverage, which may not be available. In addition, conventional V2V and V2I links may experience security issues due to malicious attacks injecting erroneous data to misguide CAVs. Finally, long trips along trans-continental highways may be stressful due to the absence of infotainment services, such as accurate navigation, news updates, and entertainment. Hence, there is an urgent need for ITS-enabling solutions in trans-continental highways.

To tackle the aforementioned challenges, we advocate the use of a HAPS constellation as a reliable provider of connectivity, caching, computing, and imaging power for ITS services. Several CAV applications can be enabled through HAPS-ITS, thus providing improved travel safety, better productivity, and greater mobility and comfort for passengers

In the remaining, HAPS-ITS refers to unmanned airships equipped with components that enable ITS services. This is motivated by the numerous advantages of unmanned airships over unmanned aircraft, including quasi-stationarity, large payload (up to 2000 kg), and extended mission times (up to 5 years).

HAPS-ITS for Improved Safety

As passenger safety is the primary concern on highways, the capacity to respond to accidents in real time increases the chance of saving lives. Indeed, passengers need to feel safe and supported in case of a hazard.

HAPS as an Accident Reporting Agent: Empowered by massive multiple-input-multiple-output (mMIMO) antennas and hybrid radio frequency (RF) and free-space optical (FSO) backhauling [6], HAPS-ITS can support emergency calls from isolated areas, and thus allows an immediate response. Also, by adopting the New Radio vehicle-to-everything (NR-V2X) communication protocol, it can guarantee V2I-supported services, such as informing vehicles of upcoming

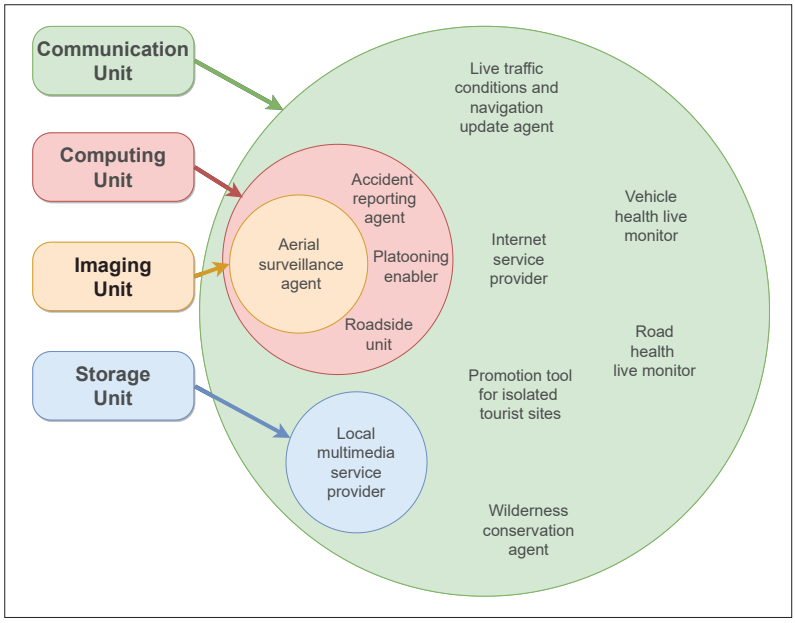


FIGURE 2. Relation between the targeted trans-continental highway services and their main enabling HAPS-ITS components (a service placed in a circle within another circle means that both units are the main drivers of that service).

hazards, monitoring their speed, and the status of roads. If an accident occurs, the HAPS system can identify it by analyzing the vehicle's data, as shown in use case (a) of Fig. 1. To assess the situation rapidly, first responders would rely on calls incoming from victims or witnesses, combined with images captured by the high-resolution cameras onboard the HAPS nodes. In fact, camera technology has improved so greatly that LiDAR cameras can constitute objects' pictures from 45 km away [7].

HAPS as a Surveillance Agent: In politically/socially unstable regions, roads crossing borders may present opportunities for criminal activities. By deploying HAPS-ITS over them, onboard cameras can monitor traffic, identify illegal activities, and respond rapidly to them. Also, data collected from CAVs through the NR HAPS-to-vehicle (H2V) links can be used to monitor vehicle speed and/or issue tickets, as illustrated in example (b) of Fig. 1. The integrity of this data can be validated by the captured video, and thus bypass any malicious data alteration. For efficient surveillance, artificial intelligence algorithms such as "You Only Learn One Representation" (YOLOR) for object detection and tracking can be called in to analyze live data, detect threats, and trigger timely responses.

HAPS as an Aerial Roadside Unit: In the future, intelligent road equipment will be intensely deployed in highways. Intelligent roads rely on IoT sensors embedded within the pavement tapes or markers. These sensors can operate with minimum maintenance using partial energy autonomy [8]. Their role includes supporting road maintenance, dynamic speed limiting for different types of vehicles, and providing information to CAVs about road conditions, weather, and traffic. To do so, they adopt communication protocols such as Wi-Fi, Bluetooth, and long-range wide area network (LoRaWAN). Along isolated highway segments, these sensors can rely on solar energy to operate, while connectivity to the control center is provided through HAPS. Due to the limited power of sensors and the high path loss, transmitting IoT data directly to HAPS may not be feasible. Alternatively, sensors can exploit passing-by CAVs as relays to forward data to the HAPS node, as shown in use case (c) of Fig. 1. Indeed, communication between sensors and CAVs can be established within a few hundred meters (e.g., using Wi-Fi or LoRaWAN) and last for a few seconds, enough to transmit IoT packets with minimal power. Moreover, V2V communications are handled by onboard transmit/receive hardware. If it fails, the vehicle cannot react to other vehicles' messages. This issue can be bypassed by forwarding V2V messages through HAPS. Despite the additional round-trip delay, to and from the HAPS node, ranging between 0.13 and 0.33 ms [4], the overall delay is still small to trigger timely actions when needed. To ensure this function, HAPS systems need to support NR-V2X communication protocols (3GPP Release 17).

HAPS-ITS for Better Productivity

As 85 percent of freight traffic travels by road, governments have invested billions of dollars in highway extensions. However, highway networks require patrolling and maintenance, often carried out with limited funding. To reduce costs, remote monitoring of vehicles/roads, and efficiently slowing or rerouting traffic can be leveraged by ITS services. The latter require V2I communications, which may not be available. HAPS-ITS can be deployed as an alternative to support V2I and provides the necessary computation power for such services. As illustrated in example (c) of Fig. 1, through the collection and analysis of the pavement sensors data, HAPS-ITS can monitor highway segments, identify critical spots, and alert the control center.

Also, platooning has been proven efficient to coordinate traffic, reduce congestion, and cut carbon footprint. It is even necessary when traveling along trans-continental highways for thousands of kilometers. Typically, it relies on V2V to exchange messages in the convoy, including vehicle accelerating/decelerating, braking, and so on. However, if a communication failure occurs, the convoy may lose formation and platooning benefits. V2I connectivity through HAPS-ITS not only allows a platoon to bypass failed V2V links, but also to leverage platooning for vehicles not yet in V2V range, as demonstrated in use case (d) of Fig. 1. Indeed, with the HAPS-ITS's wide view of the highway, it can incentivize platooning and organize convoys.

HAPS-ITS for Higher Mobility

Providing reliable and fast Internet coverage along highways has several mobility advantages. First, it increases the visibility and popularity of isolated and unknown tourist sites. Thus, local economy is positively stimulated, new businesses launched, and number of citizens permanently living in the area increased. In contrast to terrestrial BSs, a HAPS network provides connectivity without affecting the wilderness and beauty of the natural environment. This is important since several strict regulations have been put in place to preserve important sites and parks from visual pollution [9]. Finally, as highway traffic increases, HAPS-ITS processing power enables efficient traffic management through, for instance, live updates of VMS and CAVs' navigation systems.

HAPS-ITS for Enhanced Comfort

Comfort of passengers in long trips along trans-continental highways is an important issue. Whether a vehicle is autonomous or connected, passengers appreciate quickly reaching their destination. This can be enabled through HAPS-ITS by transmitting real-time traffic condition updates and rerouting information. If the vehicle is autonomous, neither the driver nor the passengers need to watch the road, as driving is monitored by the vehicle's computer and supported by the HAPS-ITS' V2I information. In such a case, passengers can use the Internet to surf, play games, or watch videos, as shown in example (e) of Fig 1. To reduce data traffic through the HAPS-ITS backhaul link and improve the traveling experience, the former can use its high caching/processing power and H2H links to act as a local multimedia server and provide vehicles with requested applications and services.

In Fig. 2, we summarize the relation between the targeted ITS services and the HAPS-ITS components.

HAPS-ITS Requirements

Since different levels of technology may be integrated into different CAVs, we address the characteristics of HAPS-ITS required to support the following CAV technology levels [10]:

- *Level 1*: CAVs equipped with driver assistance features (e.g., adaptive cruise control and lane keep assistance).
- *Level 2*: Partially automated CAVs, which assist in controlling speed, for instance, by maintaining distances from other vehicles in stop-and-go traffic and steering to center the vehicle within the lane.
- *Level 3*: Conditional automated CAVs are an upgrade of level 2 CAVs with autonomous operation under ideal conditions and limitations only (e.g., limited-access divided highways at a given speed). A human driver is still needed to take over if driving conditions fall below ideal.
- *Level 4*: Highly automated CAVs can operate by themselves, without human drivers, but they are restricted to known cases only (e.g., an autonomous bus itinerary).
- *Level 5*: Fully automated CAVs are true driveless vehicles, capable of monitoring the environment and maneuvering through any road conditions.

As shown in Fig. 2, the onboard components involved in HAPS-ITS services are the communication unit, the storage unit, the computing unit, and the imaging unit. We describe them as follows.

Communication Unit

The *communication unit* is responsible for supporting different types of communication links, namely H2H, H2V, vehicle-to-HAPS (V2H), and HAPS-to-gateway (H2G). High throughput H2H and H2G can be supported through millimeter-wave (mmWave) or FSO technologies [6, 11]. For H2V/V2H, the throughput requirement may vary in accordance with the CAV level. Indeed, we estimate the CAV data generation rate to be [10, 12]:

- *Level 1*: Between 20.05 MB/s and 40.5 MB/s
- *Level 2*: Between 120.07 MB/s and 240.7 MB/s
- *Level 3*: Between 150.09 MB/s and 350.9 MB/s
- *Level 4*: Between 160.13 MB/s and 421.3 MB/s

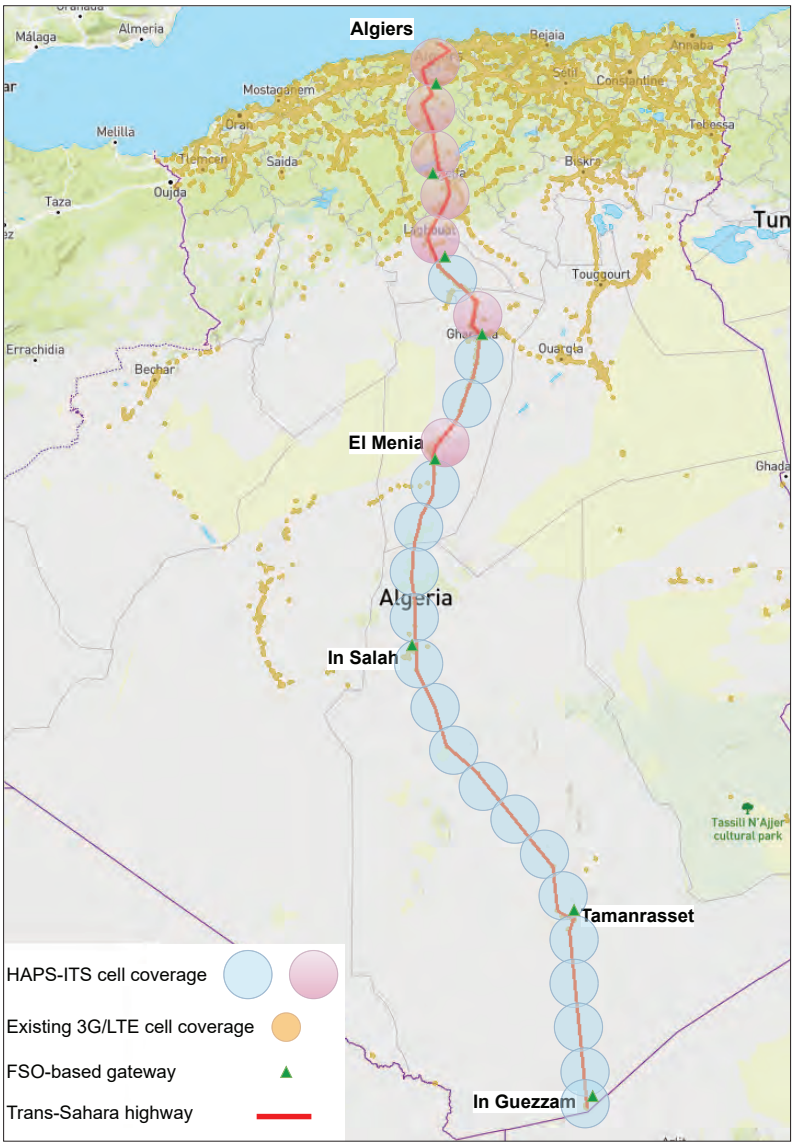


FIGURE 3. HAPS-ITS deployment along Algeria's Trans-Sahara highway (blue+red cells to support ITS services without the terrestrial network; blue cells only to support ITS services in collaboration with the terrestrial network).

- *Level 5*: Between 181.7 MB/s and 561.7 MB/s

Subsequently, only part of this data needs to be uploaded to the HAPS-ITS when traveling. Given that most of the traffic data processing is executed locally by the CAV's onboard computer and due to several security and privacy concerns, we can realistically assume that only 10 percent of this data is relevant to the HAPS-ITS. Hence, this would be equivalent to transmitting at data rates below 32, 192, 281, 337, and 450 Mb/s for CAV levels 1, 2, 3, 4, and 5, respectively. This may be achieved by HAPS-mounted large phased array antenna (i.e., mMIMO), which would provide connectivity through narrow beams and using MIMO antennas at the CAVs. HAPS-ITS can provide several services, including infotainment, fleet management and maintenance, as well as operating systems and applications. Finally, safety services require a vehicle response time below 200 ms (response time of a professional driver). The HAPS-ITS can guarantee this requirement due to its short communication delay (below 0.33 ms) [4]. However, achieving high H2V throughput is challenging as current wireless technologies, namely

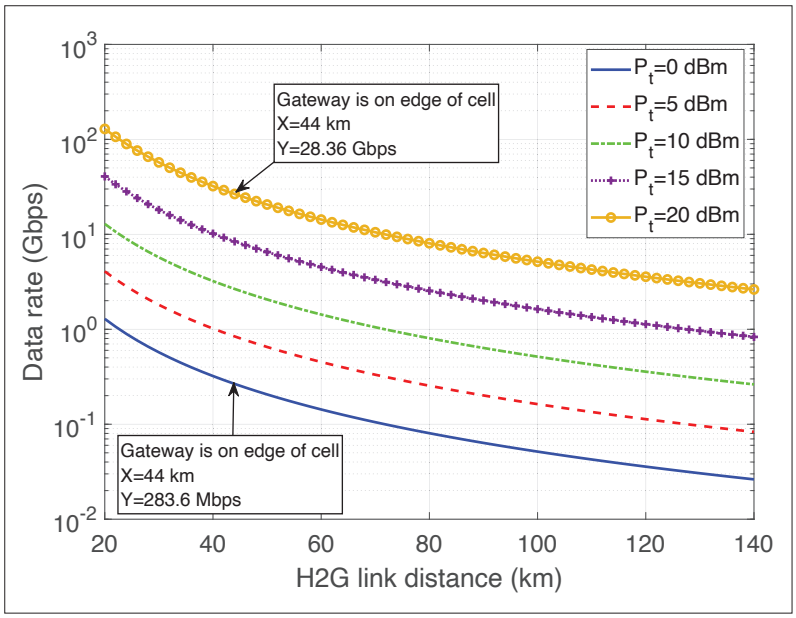


FIGURE 4. Data rate of FSO-based H2G link.

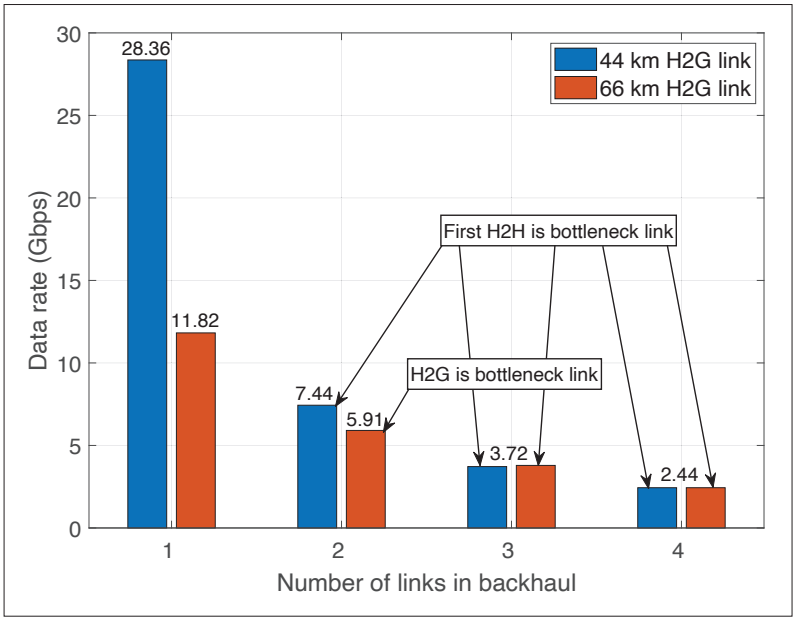


FIGURE 5. Data rate of FSO-based multi-hop backhaul (P_t = 20 dBm).

LTE, mmWave, and mMIMO, realize cell data rates under 500 Mb/s at distances up to 40 km, which may handle the traffic of low-level CAVs or of a small number of high-level CAVs simultaneously. For denser traffic along trans-continental highways, the H2V link may rapidly become the HAPS-ITS bottleneck, which would limit the efficiency of the ITS services. Finally, communication functions require a substantial number of components, such as antennas, transponders, low-noise power amplifiers, frequency converters, and filters. Since one of the HAPS-ITS roles is similar to a BS, more active components may be needed, which would increase the power consumption and require more payload space in the HAPS nodes [3].

Storage Unit

It is expected that CAVs will be equipped with *storage units* between 2 TB and 11 TB, depending on the automation level [12]. Hence, we expect the HAPS-ITS storage to range between 10 TB and 100 TB in order to handle different CAV automation levels for trans-continental highways, besides providing other services such as infotainment and fleet management. For instance, a level 1 vehicle would generate data at a rate of 145.8 GB every hour, while a level 5 vehicle's data would occupy up to 2 TB per hour. Equipping a HAPS-ITS with large storage would ensure that it could occasionally handle operations for several vehicles (e.g., accident reporting and platooning) in the event that the latter's storage and/or processing equipment fails.

Computing Unit

In CAVs, processing units collect radar and/or LiDAR and/or camera data and analyze them to provide ITS functions that range from driving assistance (level 1) to full driving automation (level 5). Thus, the computing requirements vary depending on the CAV level. For levels 1 and 2, anywhere from a few dozen to a few hundred million operations per second (MOPS) would be required to provide minimum safety assistance to drivers and infotainment services. These computing requirements can be met, for instance, with Intel processors having 32-bit microcontroller units. As CAV levels 3–5 involve features that approach full autonomous driving, more powerful computing components are required. In this matter, NVIDIA has been leading the industry with its NVIDIA DRIVE AGX autonomous vehicle solution. NVIDIA DRIVE AGX Pegasus is the most powerful system-on-chip (SoC), capable of executing 320 trillion operations per second (TOPS), thus outperforming by 10 times the NVIDIA Jetson AGX Xavier solution.[1] The computing unit of a HAPS-ITS must occasionally handle intensive tasks for a few CAVs at the same time. In other words, the HAPS-ITS can be equipped with one NVIDIA Jetson AGX Xavier to handle level 1 and 2 operations. However, at least two NVIDIA DRIVE AGX Pegasus SoCs would be required to support level 3–5 CAVs in trans-continental highways.

Imaging Unit

High-resolution cameras are an interesting add-on to the HAPS-ITS in order to monitor critical trans-continental highway segments, rapidly assess incidents, and handle inaccuracies in collected IoT data. Several technologies can be used for imaging, including optical cameras, radars, and LiDAR, which are now able to distinguish objects at very large distances [7]. These components can be installed in the HAPS-ITS to increase surveillance efficiency.

A Case Study

In this section, we simulate a HAPS-ITS deployment to provide ITS services for the Trans-Sahara highway from Algiers to Lagos (4504 km). The choice of this highway is motivated by the fact that it crosses three countries, namely, Algeria, Niger, and Nigeria, and about 70 percent of its route runs through the Sahara. For the sake of simplicity, results aree shown for the Trans-Sahara highway portion of Algeria only.

Assuming the HAPS-ITS nodes are deployed with the typical coverage footprint of 40 km in radius [3],

[1] TOPS = 10^6 MOPS.

approximately 59 HAPS-ITS nodes are required to provision ITS services for the Trans-Sahara highway, among which 26 are deployed in Algeria, as shown in Fig. 3. This number can be reduced if the existing cellular network along the highway supports ITS services. Indeed, 48 (resp. 19) HAPS-ITS nodes could fill the coverage gaps along isolated segments of the highway (resp. of the Algeria highway portion) if the existing 3G/LTE networks[2] provision ITS services, as illustrated in Fig. 3.

To support HAPS-ITS operations, a number of gateways are deployed. In Fig. 3, we see the deployment plan of FSO-based gateways (green triangles) in Algeria. Practically, only eight gateways can be deployed in cities to provide backhaul links to the HAPS-ITS constellation. Thus, several HAPS-ITS nodes along the El Menia-In Guezzam corridor rely on H2H links for backhauling.

Figure 4 shows the data rate performance of the H2G FSO link as a function of its distance for different transmit power values P_t. Assumptions, parameter values, and calculation of the data rate are conducted according to the equations within [6] for the clear sky scenario. As we see, the data rate ranges from an order of dozens to over 100 Gb/s for $P_t > 0$ dBm and H2G distance below 40 km. Practically, a powerful (resp. weak) H2G link with length 44 km would simultaneously support ITS services for 886 level 1, 147 level 2, 100 level 3, 84 level 4, and 63 level 5 (resp. 8 level 1, and 1 level 2) CAVs, respectively. As the distance increases, the FSO data rate degrades rapidly. Hence, operating H2G with low transmit power or supporting a very remote HAPS-ITS node directly through a H2G link may not be feasible, in particular for level 3–5 CAVs.

Figure 5 presents the data rate performance for FSO-based multihop backhaul links, where backhauling is modeled as a mesh communication network [13], and successive HAPS-ITS nodes are separated horizontally by an 80 km distance (corresponding to the diameter of one HAPS-ITS node footprint). For the H2H link, we adopt the assumptions and equations of [14], where FSO operates with wavelength 1550 nm and with the clear sky environment parameters as in Fig. 4. The data rate of the backhaul is dictated by the bottleneck link, defined as the minimum ratio of the link's capacity to its traffic load [13]. We found that the data rate degrades as the number of H2H links increases due to a higher traffic load coming from the supported HAPS-ITS cells along the backhaul link. Nevertheless, multihop backhauling still achieves better performance than the direct H2G link. For instance, the 121 km H2G link achieves a data rate of 3.5 Gb/s (yellow curve in Fig. 4), whereas an equivalent two-hop backhaul, with a 44 km H2G+80 km H2H, achieves data rate of 7.44 Gb/s. Finally, we notice that the H2G link length impacts the bottleneck.

In Fig. 6, we illustrate the data rate of the V2H and H2V/H2G links for different bandwidth values B when conducted through mmWave. Our study is based on the 3GPP HAPS documentation, which was detailed in [15, Subsection III-C.2]. Typically, transmit powers of HAPS and CAV are 33 dBm (3GPP TR38.811) and 24 dBm (3GPP TR36.942), respectively. Also, we assume that antenna gains are 0 dBi except for the transmit gain of HAPS equal to 43.2 dBi [15]. We

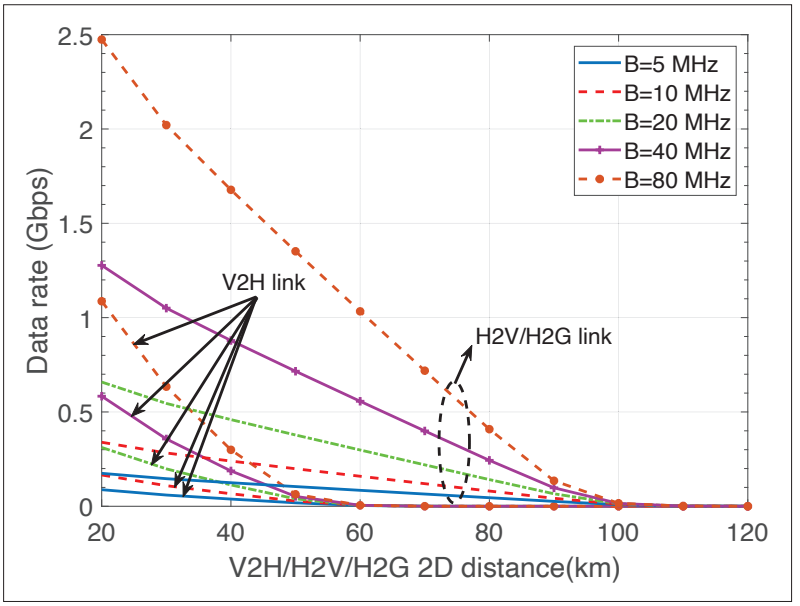


FIGURE 6. Data rate of mmWave V2H/H2V/H2G links.

set the operating frequency to f = 30 GHz and consider rural environment conditions. Clearly, the mmWave H2G is less efficient than the FSO-based one, while the H2V achieves decent data rates with high bandwidth to support ITS services. For instance, at a distance of 40 km with B = 80 MHz, the H2V link provides a downlink of 1.7 Gb/s, which can support up to four level 5 CAVs simultaneously. Moreover, V2H performance is adequate for level 3–5 CAVs only with a large bandwidth (above 20 MHz) and for a relatively short distance (below 30 km).

Challenges and Future Directions

CAVs' safety must be always guaranteed. This condition is translated into real-time and almost 100 percent reliable data sensing and transmission requirements. Part of the solution consists of organizing and configuring a constellation of HAPS nodes into a redundant real-time architecture. Moreover, current HAPS communication systems achieve reliability up to 99.9 percent, which may be tolerable for level 1–3 CAVs, but still unsatisfying for level 4 and 5 CAVs, which need reliability above 99.999 percent. In that sense, new improvements and smarter antenna technologies must be developed.

Toward communication reliability, the use of hybrid FSO/RF may be leveraged for H2H and H2G communications. Although FSO and mmWave can be complementary, a careful switching or combining design must be realized to bypass their sensitivity to weather conditions such as heavy rain, aerosols, and fog. Moreover, orthogonal access techniques for H2H and H2G links can be leveraged to avoid interference. However, interference can be experienced by ground users due to inaccurate beamforming, overlapping at the edge of HAPS cells, and dense HAPS constellations. Hence, more sophisticated access techniques and inter-HAPS coordination is needed to mitigate interference. Candidate techniques may include power control, antenna design, coordinated multipoint transmission, non-orthogonal multiple access, and mMIMO.

[2] www.gsma.com/coverage

Moreover, as work on HAPS design is ongoing, there is no clear consensus on what the lifetime of HAPS will be. Within the literature, energy consumption of HAPS-related communication functionalities have been well investigated. However, since HAPS-ITS is envisioned to support advanced applications, requiring caching, computing, and imaging, payload type and energy consumption requirements need to be further discussed. To sustain long-term HAPS-ITS operations, more energy sources have to be explored (e.g., remote charging and nuclear energy). Also, as HAPS nodes may experience failures, say, due to energy shortage, alternative approaches must be developed to continuously sustain the operations of CAVs. Potential solutions consist of the development of seamless redeployment strategies of HAPS nodes, and switching toward ground or satellite networks if available, while accepting potential degradation of services. Finally, since low Earth orbit (LEO) satellite mega-constellations are emerging to provide broadband Internet access across the globe (e.g., Starlink, OneWeb, Kuiper, and Telesat), the vertical integration of HAPS to LEO satellite networks for backhauling is necessary to achieve super connectivity [3]. However, this integration is fragile as the HAPS network's performance heavily depends on the availability of satellite links, thus making HAPS system design more complex.

Conclusion

In this article, we highlight the potential of HAPS systems for supporting ITS services for trans-continental highways. Although ITS advancements have primarily been realized within vehicles, several crucial services require V2I communications, such as the assessment of road conditions, surveillance, platooning, and recovery in case of CAV failure. These services can be supported by a HAPS-ITS network located anywhere, but particularly in isolated areas where safety is a main concern. This is demonstrated through a case study, where a practical HAPS-ITS deployment design is presented and analyzed from a communications perspective.

Acknowledgments

This work is funded by Huawei Canada and the National Science and Engineering Research Council of Canada. The authors thank Dr. Gamini Senarath, Huawei Canada, for valuable comments.

References

[1] B. Liu *et al.*, "Reducing Greenhouse Effects via Fuel Consumption-Aware Variable Speed Limit (FCVSL)," *IEEE Trans. Vehic. Tech.*, vol. 61, no. 1, Jan. 2012, pp. 111–22.
[2] C. H. Fleming and N. G. Leveson, "Early Concept Development and Safety Analysis of Future Transportation Systems," *IEEE Trans. Intelligent Transport. Sys.*, vol. 17, no. 12, 2016, pp. 3512–23.
[3] G. K. Kurt *et al.*, "A Vision and Framework for the High Altitude Platform Station (HAPS) Networks of the Future," *IEEE Commun. Surv. Tuts.*, vol. 23, no. 2, 2nd qtr. 2021, pp. 729–79.
[4] M. S. Alam *et al.*, "High Altitude Platform Station Based Super Macro Base Station Constellations," *IEEE Commun. Mag.*, vol. 59, no. 1, Jan. 2021, pp. 103–09.
[5] L. Zhu *et al.*, "Big Data Analytics in Intelligent Transportation Systems: A Survey," *Ieee Trans. Intelligent Transport. Sys.*, vol. 20, no. 1, Jan. 2019, pp. 383–98.
[6] M. Alzenad *et al.*, "FSO-Based Vertical Backhaul/Fronthaul Framework for 5G+ Wireless Networks," *IEEE Commun. Mag.*, vol. 56, no. 1, Jan. 2018, pp. 218–24.
[7] Z.-P. Li *et al.*, "Single-Photon Computational 3D Imaging at 45 km," *Photon. Res.*, vol. 8, no. 9, Sept. 2020, pp. 1532–40.
[8] S. Verma *et al.*, "Dual Sink-Based Optimized Sensing for Intelligent Transportation Systems," *IEEE Sensors J.*, vol. 21, no. 14, July 2021, pp. 15,867–74.
[9] Council of Europe, *Landscape Dimensions — Reflections and Proposals for the Implementation of the European Landscape Convention*, 2017.
[10] L. Liu *et al.*, "Computing Systems for Autonomous Driving: State of the Art and Challenges," *IEEE IoT J.*, vol. 8, no. 8, Apr. 2021, pp. 6469–86.
[11] R. Taori and A. Sridharan, "Point-to-Multipoint In-Band mmWave Backhaul for 5G Networks," *IEEE Commun. Mag.*, vol. 53, no. 1, Jan. 2015, pp. 195–201.
[12] N. Shah, B. Wang, and A. Madhok, "Storage Capacity Requirement for Autonomous Vehicles to Balloon Over 2TB in the Next Decade," *Counterpoint Technology Market Research*, Tech. Rep., 2019.
[13] W. Jaafar, W. Ajib, and S. Tabbane, "The Capacity of MIMO-Based Wireless Mesh Networks," *Proc. IEEE Int'l. Conf. Net.*, 2007, pp. 259–64.
[14] F. Fidler *et al.*, "Optical Communications for High-Altitude Platforms," *IEEE J. Sel. Topics Quantum Electron.*, vol. 16, no. 5, Sept.–Oct. 2010, pp. 1058–70.
[15] S. Alfattani *et al.*, "Link Budget Analysis for Reconfigurable Smart Surfaces in Aerial Platforms," *IEEE Open J. Commun. Soc.*, vol. 2, Aug. 2021, pp. 1980–95.

Biographies

Wael Jaafar [SM] (waeljaafar@sce.carleton.ca) is an NSERC postdoctoral fellow in the SCE Department at Carleton University, Canada. His research interests include aerial networks, 5G and beyond technologies, and machine learning.

Halim Yanikomeroglu [F] (halim@sce.carleton.ca) is a professor at Carleton University. His research interests cover 5G+ wireless networks. He is a Fellow of the Engineering Institute of Canada (EIC) and the Canadian Academy of Engineering (CAE), and he is a Distinguished Speaker for IEEE Communications Society and IEEE Vehicular Technology Society.

CALL FOR PAPERS

IEEE COMMUNICATIONS MAGAZINE

FEATURE TOPIC: AFFECTIVE COMPUTING AND COMMUNICATIONS RESEARCH, TECHNOLOGIES, MANAGEMENT, AND APPLICATIONS

BACKGROUND

Despite great advances in communications technologies and applications, considerable aspects of face-to-face human interactions remain difficult to relay. Instances such as the introduction of emojis and the proliferation of video-based communications have certainly enriched our communications. Still, much needs to be sought in both human-to-human as well as human-to-machine communications.

A computer's ability to capture a human's emotion is called affective sensing, alternatively called emotion sensing and more recently emotional artificial intelligence (emotional AI). Since its formulating introduction in mid 1990s by Picard, affective sensing has evolved substantially, and have found its role in various applications. These range from understanding the driver's affective state, distraction and behavior; to automating the capture of user's Quality of Experience (QoE) for a certain connection; to the use of emotions in crowd and traffic management; to relaying emotions to and from the metaverse.

This Feature Topic addresses affective computing and communications through research advances, technologies, applications, and management aspects that enable human-computer interaction including emotion sensing. The goal is to report on the most up-to-date contributions in this area. Affective computing and communications must be central to all topics that include, but are not limited to the following:

Areas of interest include the following.

- Application of emotion sensing in communications, including but not limited to
  - Human-to-human communications
  - Human-to-machine communications
  - Emotions in social networks communications
  - Metaverse and mixed-reality (XR) environments
  - eHealth, mHealth
  - Education
  - Video games
  - Network management, automated QoE capture
- Connected vehicles, telematics, driver behavior monitoring
- Advances in emotion sensing, modelling, and discretization.
- Emotion classification, taxonomy
- Detecting and recognizing emotional information
- Fusion considerations in emotion sensing
- Reports on implementations and testbeds of emotion-enabled communications
- Architectural requirements for emotion-enabled communications.
- The use of artificial intelligence and machine learning in emotion sensing and emotion-enabled communications.
- Considerations for cross-layer communications and network management.
- Validation requirements for emotion-enabled communications.
- Communication Technologies for emotion sensing
- Emotion sensing for Fifth Generation (5G) and Sixth Generation (6G) networks
- Tactile Internet and affective computing
- Haptic Computing and communications

SUBMISSION GUIDELINES

Manuscripts should conform to the standard format as indicated in the Information for Authors section of the *IEEE Communications Magazine's* Manuscript Submission Guidelines. Please, check these guidelines carefully before submitting since submissions not complying with them will be administratively rejected without review.

All manuscripts to be considered for publication must be submitted by the deadline through Manuscript Central. Select the FT-2223/Affective Computing and Communications" topic from the drop-down menu of Topic/Series titles. Please observe the dates specified here below noting that there will be no extension of submission deadline.

IMPORTANT DATES

Manuscript Submission Deadline: 1 December 2022
Decision Notification: 1 March 2023
Final Manuscript Due: 15 April 2023
Publication Date: Third Quarter 2023

GUEST EDITORS

Abd-Elhamid M. Taha
Alfaisal University, Kingdom of Saudi Arabia
ataha@alfaisal.edu

Ajjen Joshi
Affectiva, USA
ajjen.joshi@smarteye.ai

M. Fatima Domingues
Instituto de Telecomunicações-Aveiro, Portugal
fatima.domingues@ua.pt

Mehmet Ulema
Manhattan College, USA
m.ulema@ieee.org

ACCEPTED FROM OPEN CALL

Engineered Electromagnetic Metasurfaces in Wireless Communications: Applications, Research Frontiers and Future Directions

Mohsen Khalily, Okan Yurduseven, Tie Jun Cui, Yang Hao, and George V. Eleftheriades

This tutorial is aimed at presenting a plethora of EM metasurface applications and research frontiers to illustrate a broader impact of SEM in real-world platforms, advanced communication systems and new devices.

Abstract

Surface electromagnetics (SEM), as a sub-discipline of electromagnetic (EM) science, is strongly linked with the ability to manipulate an arbitrary EM wavefront. This exceptional capability of manipulating the surface-bound and free-space EM waves for the guidance and control of anomalous reflection, refraction and transmission has catapulted an abundance of new research frontiers. This has resulted in the realization of many novel applications for modern real-life platforms and the introduction of several new modelling techniques and engineering approaches to give rise to some unconventional devices. Consequently, EM engineered metasurfaces, due to numerous emerging applications, are beginning to revolutionize the EM industry. Recently, the practical usage of metasurfaces has gained a substantial amount of interest and traction for a wide range of applications in microwave, millimetre-wave (mmWave), Terahertz (THz) and even optical wavelengths. This tutorial is aimed at presenting a plethora of EM metasurface applications and research frontiers to illustrate a broader impact of SEM in real-world platforms, advanced communication systems and new devices.

Introduction

Recent advances in surface electromagnetics (SEM) in the context of metasurface engineering have opened new opportunities in EM aperture design. A metasurface aperture is made up of elements designed at sub-wavelength scale that allow for SEM properties to be manipulated. This sub-wavelength aperture control of amplitude and phase distributions is the key enabling factor in achieving the desired beam direction, shape, and polarization [1]. As an example, metasurface antennas can exhibit a compact form-factor and overcome the existing limitations of large and heavy conventional antenna modalities such as in the case of parabolic reflector antennas, lens antennas, transmit-arrays, and reflect-arrays. They also offer significant advantages over phased array-based modalities that can suffer from high cost and high losses that grow substantially with aperture size and frequency. These advantages include eliminating the need for dedicated phase shifters and power amplifiers, and can drastically simplify the hardware architecture.

In this article, we present the potential of metasurfaces for emerging wireless networks and applications. Specifically, the contributions of this article are as follows:

- We investigate the concept of metasurfaces and explore their emerging state-of-the-art applications to existing microwave, millimeter-wave (mmWave), terahertz (THz), and optical wireless systems.
- We identify the potential of metasurfaces and their fundamental challenges associated with SEM in these areas.
- We propose future research directions and opportunities to tackle these challenges with the steps to be taken in order for metasurface apertures to be a realistic enabling technology for future wireless communication systems.

The ultimate vision of this article is to identify opportunities in SEM, and present the impact of metasurface apertures, their challenges, and their future potential as an enabling physical layer technology in transceiver design for wireless systems and applications. The outline of this article is as follows. After briefly introducing the concept of metasurface antennas, we present the application of EM metasurfaces in 6G communication technology and beyond, including within the THz band spectrum and emphasizing on reconfigurable intelligent surfaces (RISs). Next, we discuss the metasurface concept in the context of imaging and direction of arrival (DoA) estimation, which is of particular interest for channel characterization in wireless communications. Following, we explore the application of the metasurface concept to satellite communications. Finally, we present a discussion on SEM in optical communications, provide future research trends of EM metasurfaces, and present the concluding remarks.

Metasurface Antennas

Metasurface antennas are synthesized using periodically arranged sub-wavelength elements (unit cells) conventionally printed on a piece of dielectric material [2, 3] to manipulate the incident wavefront with sub-wavelength spatial resolution. Similar to an optical hologram, holographic beamforming at microwave and mmWave frequencies can be achieved by treating the incident wave (i.e., the EM source exciting the metasurface) as a reference wave, and the desired radiation pattern as the objective func-

Digital Object Identifier:
10.1109/MCOM.004.2200052

0163-6804/22/$25.00 © 2022 IEEE

tion. In this process, the role of the metasurface is to act as a hologram, converting the reference-wave to a desired field distribution at the antenna aperture, radiating the far-field radiation pattern of interest as depicted in Fig.1. In other words, when the metasurface calculated by interacting the reference-wave and the objective function is excited by the reference-wave, it is ensured that the metasurface radiates the desired radiation pattern. As a consequence of this definition, it can be concluded that a metasurface can be manipulated by encoding it with the necessary phase information to provide a certain functionality. Holographic surfaces are designed in such a way that the source field is tuned to radiate in a certain manner by guiding waves toward a radiating aperture or by supporting leaky waves that exhibit the desired radiation characteristics.

Although metasurface antennas hold great potential in addressing size constraints and system complexity of the conventional antenna modalities, certain challenges still exist that need to be addressed before they can be deployed commercially. Unit cells, which form the metasurface's aperture, conventionally exhibit a strong frequency dependency, and therefore metasurfaces are highly frequency dispersive in nature. This reduces the achievable operating bandwidth of metasurface antennas. Furthermore, electrically tunable and reconfigurable antennas with high gain are required to satisfy various targeted applications. These requirements can limit the metasurface antennas' performance in terms of losses, aperture efficiency, operational bandwidth, and complexity, especially when the size of the antenna and the number of elements have to be increased while still keeping the scale of the elements at sub-wavelength.

The operational bandwidth of metasurface antennas can be increased by manipulating the SEM properties in both the spatial and spectral domains. This can be achieved by utilizing spatial field tuning offered by metasurfaces jointly with existing filtering mechanisms that have proven to be successful in the microwave and mmWave bands. Furthermore, advanced methods that employ control circuits within metasurfaces can address the problems associated with reconfigurability. Particularly interesting in this context is that spatial light modulators employing liquid crystal on silicon technology can be used to realize electrically addressable unit cells; however, such advanced circuitry is currently not available at microwave and mmWave frequencies. Such circuitry can also address the problem of synthesis for multiple-input multiple-output (MIMO) device operation where the beam is constructed by combining the modified beams of multiple cascaded metasurfaces, thereby enhancing the overall performance and allowing for a greater number of functions to be embedded within the device.

Application of Metasurfaces in 6G Communications and Beyond

Metasurfaces for Coverage Enhancement

Metasurface apertures have recently received significant interest in 6G and beyond applications, particularly in the context of RIS apertures. An RIS is a large metasurface backed by a control unit and synthesizes a radiation pattern of interest using either holography theory or generalized Snell's law. The role of the RIS is to modulate the incident wave impinging on the aperture into a desired aperture field that gives rise to a radiation pattern of interest. An RIS aperture can realize this conversion at a sub-wavelength level by altering the phase of each unit cell. As a result, an RIS has a flat-panel topology and still offer the ability to reconfigure the radiation pattern in a purely electronic manner without the need for dedicated phase shifting circuits. This all-electronic operation also eliminates the need for mechanical scanning to steer the radiation pattern of RISs, thereby making them a serious candidate for 6G and beyond applications.

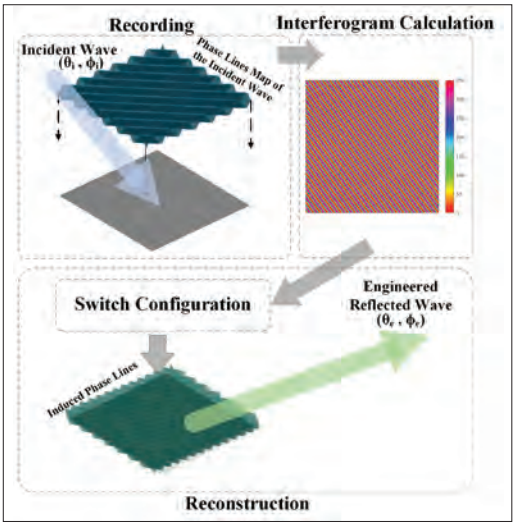


FIGURE 1. Holographic beam-synthesis process. A metasurface is modulating the reference-wave to the desired radiation pattern.

RIS layout can be designed based on two different concepts: continuous phase gradient (analog) metasurface [4] or digital coding metasurface [5]. Analog surfaces can provide more freedom to tune the phase values and require switches to manipulate the radiation pattern. A continuous phase gradient-based RIS has been developed at 6GIC, University of Surrey, and the prototype is illustrated in Fig. 2. Using this RIS, a software-defined radio (SDR) system was employed to stream a video signal over the carrier frequency of 3.5 GHz using quadrature phase shift keying (QPSK) modulation with a forward error correction (FEC) rate of 7/8 [4]. At the receiver side, without the RIS, the signal level was lower than the minimum required sensitivity level for the video stream to be decoded. When the developed RIS is introduced to the propagation environment, it provides sufficient gain toward the receiver where the signal level was observed with 15 dB enhancement compared to the case when the RIS was powered off.

Digital coding and programmable metasurfaces can realize a large number of distinct functions by encoding different digital states of unit cells, and switch these functions in real time by using a field programmable gate array (FPGA). Because of the representation of metasurfaces by the digital coding states (Fig. 3), it is possible to bridge the EM physical world and the digital world using the metasurface platform. In this manner, the dig-

The operational bandwidth of metasurface antennas can be increased by manipulating the SEM properties in both the spatial and spectral domains. This can be achieved by utilizing spatial field tuning offered by metasurfaces jointly with existing filtering mechanisms that have proven to be successful in the microwave and mmWave bands.

The ultrawide THz frequency bands with a potential capacity of several terabits per second are predicted to enable next generation wireless communication applications, such as big-data wireless cloud, massive data in large-scale digitalization, ultra-fast wireless download (data kiosk), seamless data transfer, and large uncompressed data links for low latency.

FIGURE 2. RIS developed at 6GIC, University of Surrey.

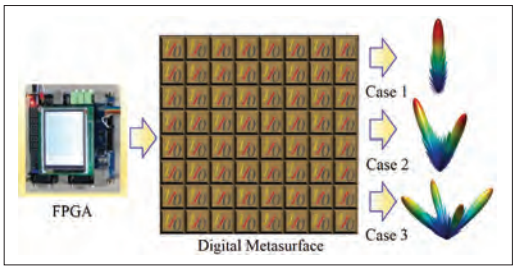


FIGURE 3. The digital coding metasurface. Different digital coding sequences of 0 and 1 states will result in different functions.

ital coding and programmable metasurfaces can process the digital information directly; hence, they are also called "information metasurfaces" [6]. Recently, time-domain coding and space-time coding metasurfaces have been introduced, which can manipulate the spatial EM beams, the frequency spectra, and digital information simultaneously and independently [6]. Hence, the information metasurfaces are no longer effective materials like the traditional metamaterials, but also are information systems. In particular, the information metasurfaces have been successfully applied to build up simplified-architecture and new-architecture wireless communication systems, such as binary frequency shift keying (BFSK), quaternary PSK (QPSK), 8PSK, and 16 quadrature amplitude modulation (16QAM), in which the information modulation process is directly performed on the metasurface interface without using the digital-analog conversion and mixing process [6]. In addition to the reflective topology, RIS architectures simultaneously transmitting and reflecting (STAR) have also received significant attention recently to realize both reflecting and transmitting modes to manipulate EM waves [7]. In the STAR implementation, a part of the signal is reflected to the reflection space, while the remaining part is radiated into the transmission space, thus facilitating the full-space manipulation of signal propagation.

In the future, the combination of new-architecture wireless systems and new channel characteristics will produce revolutionary developments in wireless communications and networks.

A significant hurdle for considering RIS as a feasible solution for wireless communications is the difficulty in estimating the channel characteristics on a real-time basis. The compressive sensing-based techniques, as outlined later, can offer a promising potential to achieve DoA estimation from metasurface-based apertures, thereby enabling RISs to retrieve the DoA estimation of the end user [8].

Application of Metasurfaces in THz Communications

The ultrawide THz frequency bands with a potential capacity of several terabits per second are predicted to enable next generation wireless communication applications, such as big data wireless cloud, massive data in large-scale digitalization, ultra-fast wireless download (data kiosk), seamless data transfer, and large uncompressed data links for low latency. In order to facilitate efficient and compact THz devices, strong interactions are required between THz radiation and materials. Conventional materials with magnetic response in the THz regime are particularly rare. In this regard, the remarkably strong EM responses offered by THz metasurfaces with loss suppression and exotic EM properties (e.g., negative index media) can lead us to develop an entire class of THz devices. Modulators with high modulation depth and fast temporal responses are crucial for THz systems with terabits-per-second data rates, for wireless communications, imaging, and sensing. Metasurfaces are a potential candidate for compact and efficient THz modulators due to their sharp response in both reflection and transmission modes as well as they dramatically enhance the field confinement within the unit cells. Besides the modulator-related research, the amplitude and phase control capability of metasurfaces enables the arbitrary manipulations of THz radiation, including highly efficient broadband polarization conversion, broadband anomalous reflection with strong phase discontinuity, and simultaneously controllable phase, amplitude, and polarization in multiple bands.

Another emerging SEM-based engineering application is the metasurface absorbers with the capability of achieving near-unity absorption. The phenomenon of completely absorbing the impinging THz waves on metasurfaces can be obtained by manipulation of both the electric and magnetic properties. THz metasurface absorbers can greatly improve the energy management in wireless networks and facilitate the design of THz sensors with high sensitivity [9]. A further SEM-based technology that can be considered for future wireless communications systems is RIS as outlined in this article. However, RISs at these high frequencies play an outstanding role compared to the microwave and mmWave regimes. As THz metasurfaces allow for the manipulation of the phase, amplitude, and polarization state of THz waves, non-line-of-sight (NLoS) communications can be enabled by redirecting the impinging waves toward the angle of interest to compensate for the significant losses caused by the blockage THz waves in a wireless communication environment.

Application of Metasurfaces in Imaging and Direction of Arrival Localization

Another emerging application of SEM is to realize compressive sensing-based techniques. In this context, a particularly exciting application of SEM in the metasurface context has been demonstrated in computational imaging and DoA localization [8]. From an EM point of view, imaging is an inverse problem involving the reconstruction of an object from the knowledge of incident and scattered fields. In this process, it is vital that a sufficient number of measurements are collected for the object being imaged to gather enough feature details so that an estimated image of it can be recovered. Conventionally, imaging requires raster scanning the scene, either mechanically or electronically or by a combination of both. The two most widely used techniques that can be considered in this context are synthetic aperture radar (SAR) and phased arrays. Traditionally, SAR is mostly adopted to raster scan the scene by mechanically scanning an antenna (or an array of antennas) to synthesize an electrically large effective aperture. The phased array technique, on the other hand, is implemented to synthesize an effective aperture by means of controlling the phase response of each antenna element within the array that collectively raster scans the scene information.

Due to the mechanical scanning requirement, data acquisition speed for the SAR technique can be extremely slow. This challenge can be addressed using the all-electronic phased array technique; however, as covered earlier, phased arrays require a dedicated phase shifting circuit for each antenna element within the array, significantly increasing the hardware layer complexity and power consumption.

EM metasurfaces have recently been shown to offer a promising alternative to these techniques and pave the way for compressive sensing to address these challenges. An interesting example was demonstrated using a wave-chaotic metasurface antenna that radiates a spatio-temporally incoherent radiation pattern as a function of a frequency sweep — a technique also known as frequency diversity. In other words, as the operating frequency is swept within a selected frequency band, the radiation pattern of the metasurface changes on a quasi-random basis. These quasi-random radiation patterns are then used to illuminate the scene information and collect the back-scattered data. This results in two important outcomes: First, the back-scattered signal is encoded and compressed by the transfer function of the metasurface antenna. Second, the scene information is captured in an indirect manner — in other words, the scene is illuminated using quasi-random patterns replacing the pixel-by-pixel raster scanning of conventional SAR and phased array techniques. Figure 4 depicts the compressive imaging process facilitated by a frequency-diverse metasurface antenna radiating frequency-dependent, quasi-random radiation patterns. The imaged object in the presented example consists of the word "QUB," referring to Queen's University Belfast. The quasi-random radiation patterns generated by the metasurface diverge from the pixel-by-pixel raster scanning requirement in that each radiation pattern illuminates a random subset of multiple pixels in the scene. This wave-chaotic probing of the scene information results in significant simplification in the physical hardware layer by eliminating the need for Nyquist sampling at the aperture plane, and instead compressing the back-scattered data into a single channel (i.e., single-pixel architecture) as depicted in Fig. 4.

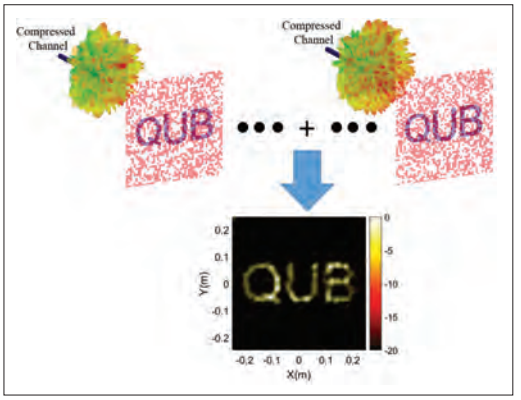


FIGURE 4. Compressive imaging using mmWave coded-aperture metasurface antennas. The colorbar is in dB scale.

The concept of compressive sensing using wave-chaotic metasurface antennas has recently been extended to wireless communications, particularly in the context of localization and DoA estimation for channel characterization [8]. A significant advantage of such compressive DoA estimation modalities is that a wave-chaotic metasurface can be used to listen to incoming EM sources and compress the received signal into a single channel, replacing the conventional array-based receiver architectures for DoA estimation. The compressive DoA modality offers significant potential for channel estimation problems, a key requirement for the physical antenna layer in a wireless communication system to synthesize the desired radiation pattern.

Application of Metasurfaces in Satellite Communications

Another interesting application of SEM-based metasurface apertures can be considered in satellite communications (satcom). Antennas for satcom applications typically require a high antenna gain, a metric that can be satisfied using electrically large aperture modalities such as the conventional parabolic reflector architecture. However, for space missions, such antennas have proven to be quite challenging to launch because they require a considerable amount of payload space and can be quite heavy [10]. Moreover, these antennas are static in nature and therefore require mechanical movement of the antenna aperture to scan the radiation pattern. A planar alternative can be considered in the context of array antennas, thereby simplifying the launching process of the satellite significantly. However, such antennas typically utilize the phased array technique for beam synthesis as they require full phase control for each individual antenna element across the array aperture. As covered earlier, the need for phase-shifting circuits not only increases the hardware complexity; it can also increase the total power consumption significantly.

From an EM point of view, imaging is an inverse problem involving the reconstruction of an object from the knowledge of incident and scattered fields. In this process, it is vital that a sufficient number of measurements are collected for the object being imaged to gather enough feature details so that an estimated image of it can be recovered.

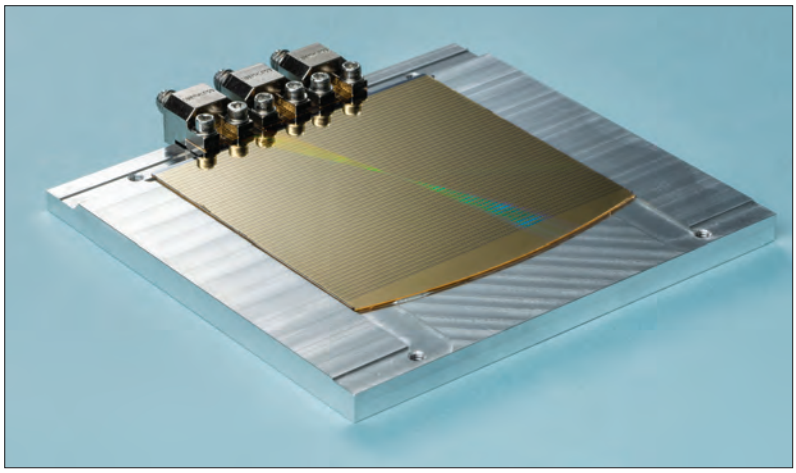


FIGURE 5. Flat-panel, reconfigurable metasurface antenna.

SEM facilitated by metasurfaces offers an alternative solution for engineering a desired wavefront at the antenna aperture. A significant advantage of this technique is that a radiation pattern of interest can be synthesized at a sub-wavelength level using an all-electronic technique without the need for phase shifting circuits. This can not only simplify the hardware architecture; it can also substantially reduce the power consumption and form-factor of the antenna. In this context, holographic metasurfaces can play a fundamental role in replacing the conventional reflector and array-based antenna architectures with flat-panel, programmable antenna architectures. Leveraging the holographic beamforming concept, several promising solutions in the context of satcom systems have been demonstrated. An example is the holographic beamforming process realized by means of designing an impedance surface that takes a known surface-wave distribution as input and modulates it into a desired aperture field distribution as output [11]. Using the holographic principle, the main objective (i.e., desired aperture field distribution) is realized by the interaction of the calculated impedance surface with the surface-wave excitation that gives rise to the radiation pattern of interest in the far-field of the antenna. In [11], this concept was used to realize a downlink channel from low Earth orbit (LEO) satellites at Ka-band frequencies, which demonstrates vastly reduced hardware complexity without sacrificing antenna performance. Such a design can offer significant potential to meet the stringent weight, space, and power requirements of satcom antennas.

At higher frequencies, the concept of holographic metasurface antenna can be quite advantageous compared to the phased array technology due to the quadratically increasing number of array elements required in phased arrays, and hence a higher number of phase shifters and power amplifiers, as the frequency increases. As an example, a semiconductor-based holographic metasurface antenna for a space-based mission was developed, shown in Fig. 5, for operation at 94 GHz within the W-band frequency range [3]. Using the holographic beamforming concept, the radiation pattern of the metasurface is synthesized by taking a known reference-wave (or guided-mode) as input interacting with the metasurface layer to realize an objective function (desired radiation pattern of interest) as depicted in Fig. 1. By applying this rather simple concept, reconfigurable beam synthesis has been demonstrated using the developed metasurface, which is only 525 microns thick and exhibits a truly flat-panel architecture. These demonstrations provide a testament to the significant potential of SEM in space-based communication systems.

A similar concept can also be used to build radome-based enclosure structures for mmWave systems with minimal reflection [12]. Despite the application versatility and advantages offered, metasurface antennas for satellite applications currently have several challenges, as summarized in Table 1. First, the achievable frequency bandwidth from such antennas can be limited, which would require additional techniques to be implemented to widen the operating frequency band of these systems. Second, the aperture efficiency of metasurface antennas tend to be lower when compared to some conventional antenna architectures, such as parabolic reflectors. Addressing these technological challenges constitutes an extremely active research field.

Application of Metasurfaces in Optical Communications

Going beyond the microwave, mmWave, and THz spectra of the EM spectrum that have been covered so far, metasurface-based aperture modalities have recently received significant attraction as a means to manipulate waves at optical frequencies. Metasurfaces such as dielectric EM surfaces (DEMSs) have shown the potential to be utilized in designing optical devices to overcome the limitations of metals at optical frequencies (e.g., high loss and dispersive behavior). DEMSs can be utilized in developing numerous optical and photonic components such as flat lenses, holograms, anomalous deflectors, beam splitters, optical cloaks, and wave plates without involving metals. Such EM surfaces are commonly employed in optical and photonic systems. A typical optical DEMS-based component is the famous beam splitter. Some DEMS-based beam splitters are developed based on the difference between the incident light's wavelengths, where the reflected or refracted beams with different frequencies are efficiently directed to different paths. Beam splitters can also be designed by utilizing the polarization characteristics of the incident light where the incident waves with different polarizations are deflected into different directions. Here, an EM surface can be created by introducing the polarization-dependent localized phase distribution. DEM-based beam splitters can play a crucial role to achieve spatial data multiplexing in optical communications and meet the increasing demands for channel capacity at optical frequencies [13].

A significant challenge in optical communication systems can be considered in the context of etendue, posing a fundamental limit on data speed that can be transmitted at optical frequencies. In order for optical frequencies to be a viable technology for wireless communications, detectors with large aperture sizes are needed, which fundamentally, reduces their response time and hence their speed. This places a significant burden on free-space optical communication

Application	Pros and cons of SEM	Pros and cons of existing solutions
6G and beyond	Pros 1: Reduced power consumption and simplified hardware layer by eliminating phase shifters (e.g., holography). Pros 2: Low form-factor. Pros 3: Capability to realize multiple EM functions from the same surface. Cons 1: Limited frequency bandwidth (additional techniques are available to widen the frequency band of operation). Cons 2: Low to moderate aperture efficiency.	Pros 1: Moderate to high aperture efficiency. Pros 2: Superior beam scanning performance. Cons 1: Increased hardware complexity. Cons 2: Increased power consumption. Cons 3: Increased hardware footprint.
Imaging and localization	Pros 1: Single-pixel, compressive imaging leading to a drastically reduced number of channels. Pros 2: All-electronic operation (no mechanical scan needed). Pros 3: Reduced power consumption. Pros 4: Simplified hardware layer. Cons 1: Increased computational complexity.	Pros 1: Minimal computational complexity. Pros 2: Low computation cost. Cons 1: High power consumption (for phased arrays). Cons 2: Mechanical scan (for conventional SAR). Cons 3: Increased hardware complexity and footprint.
Satellite communications	Pros 1: Reduced power consumption and simplified hardware layer by eliminating phase shifters (e.g., holography). Pros 2: Low form-factor — ideal for instrument integration. Pros 3: Capability to realize multiple EM functions from the same surface. Cons 1: Limited frequency bandwidth (additional techniques are available to widen the frequency band of operation). Cons 2: Low to moderate aperture efficiency.	Pros 1: Moderate to high aperture efficiencies. Pros 2: Large frequency bandwidths. Cons 1: Bulky hardware architecture and high storage volume. Cons 2: Static operation; beam-scanning requires a mechanical scan. Cons 3: Complex deployment.
Optical communications	Pros 1: Low form-factor facilitated by a flat-panel topology replacing conventional bulky lenses. Pros 2: Capability to realize multiple optical wavefronts, such as Laguerre-Gaussian and OAM, from the same surface. Cons 1: Limited frequency bandwidth (additional techniques are available to widen the frequency band of operation). Cons 2: Challenging fabrication requirements typically leading to low efficiencies.	Pros 1: Increased flexibility in fabrication requirements. Pros 2: Potential to achieve higher efficiencies. Cons 1: Bulky hardware architecture. Cons 2: Limited opportunities to address etendue-related challenges using conventional optical modalities.

TABLE 1. Pros and cons analysis of SEM-based and existing solutions for different applications.

technology, and can be considered the major reason data transmission at optical frequencies is currently restricted to fiber optic cables. Optical metasurfaces have recently shown significant potential to address this challenge [14]. To optimize the aperture size of optical sensors for wireless communications, lenses can also be essential components in optical systems. However, the conventional lenses suffer from having a bulky profile and low efficiency. Such issues could be overcome by using metalenses, particularly in the context of flat lenses, generating a phase distribution with a nano-scale resolution. By utilizing metasurface technology, this resolution is achievable as the interaction of light is localized on each unit cell of the metasurface. Thus, by mapping the required phase profile, the wavefront of light can be reshaped into a focused or defocused beam, which can achieve better performance compared to the conventional lenses.

By adopting the metasurface technology, not only can the focused/defocused and split optical beams be generated, but also, optical beams with twisted wavefronts can be realized, including Laguerre–Gaussian (helical) beams, higher-order Bessel–Gaussian beams, and so on. Among them, optical vortex beams with helix-shaped wavefronts, known as orbital angular momentum (OAM) beams, have captured quite a lot of interest as they can be utilized in high-speed free-space optical communications to provide an additional multiplexing channel [15]. The main reason behind this is that OAM beams with different numbers of twists of the wavefront contained within one wavelength cannot interfere with each other and hence can be multiplexed. Usual approaches of OAM beam generation utilize the Pancharatnam–Berry (P-B) phase. This phase profile can also be produced using metasurface technology; however, P-B phase requires circularly polarized incident light to be generated. In order to overcome this limitation of P-B phase generation, OAM can be generated by arranging unit cells with $0–2\pi$ radians phase change around the azimuthal direction, thereby utilizing a metasurface's abrupt phase shifting capability. Despite the challenges associated with fabricating sub-wavelength structures at optical frequencies, metasurface-based apertures offer significant potential for optical communications.

Future Research Trend of Engineered Electromagnetic Metasurfaces

Engineered EM metasurfaces represent a significant improvement from traditional EM aperture solutions by offering full control of the aperture fields due to their capability of accurately designing the impedance boundary conditions. While the EM metasurfaces have shown their ability to be practically used at microwave, mmWave, THz, and even optical wavelengths for different engineering applications, there is much research work to be conducted. Future research directions extend from developing reliable dynamic wavefront controllers to the design of conformal EM metasurfaces. The dynamic control of the EM wavefronts is appealing for future wireless 6G communications, particularly in beamforming and tracking applications at mmWave and THz frequency bands. Dynamically reconfigurable metasurfaces are also needed for automotive radar applications, vehichle-to-vehicle (V2V) communication, and satcom systems. Furthermore, for fast moving vehicles, conformal apertures can be used to reduce the impact of weather on reception. The limited frequency bandwidth is one of the significant limitations of EM metasurfac-

es; hence, bandwidth enhancement is another important research topic. Future modern wireless technologies are expected to support multi-gigabit transmission; hence, mmWave and THz hardware cost and efficiency will play an important role to deploy high-frequency systems. In this regard, the gap-waveguide technology can provide efficient and low-loss mmWave and THz components.

Conclusion

In this article, we provide an overview of metasurface-based aperture modalities as an enabling technology for future EM surface applications in real-world wireless platforms. The metasurface concept offers a drastically different approach to EM wavefront engineering, emerging as a disruptive technology to the existing antenna architectures. As the demands for increased data speeds, reduced power consumption, minimized device footprint, and simplified hardware architecture are becoming critical factors for next-generation microwave, mmWave, THz, and optical systems, the metasurface concept offers itself as a promising architecture to address these demanding challenges.

Acknowledgment

The work of O. Yurduseven was supported by the Leverhulme Trust under Research Leadership Award RL-2019-019.

References

[1] A. Epstein and G. V. Eleftheriades, "Huygens' Metasurfaces via the Equivalence Principle: Design And Applications," *JOSA B*, vol. 33, no. 2, 2016, pp. A31–A50.
[2] Y. Wang *et al.*, "Broadband High-Efficiency Ultrathin Metasurfaces with Simultaneous Independent Control of Transmission and Reflection Amplitudes and Phases," *IEEE Trans. Microwave Theory and Techniques*, vol. 70, no. 1, 2022, pp. 254–63.
[3] O. Yurduseven *et al.*, "Multibeam SI/GAAS Holographic Metasurface Antenna at W-Band," *IEEE Trans. Antennas and Propagation*, vol. 69, no. 6, 2021, pp. 3523–28.
[4] A. Araghi *et al.*, "Reconfigurable Intelligent Surface (RIS) in the Sub-6 Ghz Band: Design, Implementation, and Real-World Demonstration," *IEEE Access*, vol. 10, 2022, pp. 2646–55.
[5] T. J. Cui *et al.*, "Coding Metamaterials, Digital Metamaterials and Programmable Metamaterials," *Light: Science & Applications*, vol. 3, no. 10, 2014, pp. e218–e218.
[6] T. J. Cui *et al.*, "Information Metamaterial Systems," *Iscience*, vol. 23, no. 8, 2020.
[7] Y. Liu *et al.*, "Star: Simultaneous Transmission and Reflection for 360° Coverage by Intelligent Surfaces," *IEEE Wireless Commun.*, vol. 28, no. 6, Dec. 2021, pp. 102–09.
[8] O. Yurduseven *et al.*, "Frequency-Diverse Computational Direction of Arrival Estimation Technique," *Scientific Reports*, vol. 9, no. 1, 2019, pp. 1–12.
[9] S. B. Amlashi *et al.*, "Surface Electromagnetic Performance Analysis of a Graphene-Based Terahertz Sensor Using a Novel Spectroscopy Technique," *IEEE JSAC*, vol. 39, no. 6, 2021, pp. 1797–1816.
[10] N. Chahat *et al.*, "Cubesat Deployable KA-Band Mesh Reflector Antenna Development for Earth Science Missions," *IEEE Trans. Antennas and Propagation*, vol. 64, no. 6, 2016, pp. 2083–93.
[11] G. Minatti *et al.*, "Ka-Band Metasurface Antenna for Data Downlink from LEO Satellites," *12th Euro. Conf. Antennas and Propagation*, 2018.
[12] Y. He and G. V. Eleftheriades, "A Thin Double-Mesh Metamaterial Radome for Wide-Angle and Broadband Applications at Millimeter-Wave Frequencies," *IEEE Trans. Antennas and Propagation*, vol. 68, no. 3, 2020, pp. 2176–85.
[13] S. Kruk *et al.*, "Dielectric Metasurfaces for Optical Communications and Spatial Division Multiplexing," *CLEO: QELS Fundamental Science*, OSA, 2018, pp. FW3H–4.
[14] A. J. Traverso *et al.*, "Low-Loss, Centimeter-Scale Plasmonic Metasurface for Ultrafast Optoelectronics," *Optica*, vol. 8, no. 2, 2021, pp. 202–07.
[15] E. Karimi *et al.*, "Generating Optical Orbital Angular Momentum at Visible Wavelengths Using a Plasmonic Metasurface," *Light: Science & Applications*, vol. 3, no. 5, 2014, pp. e167–e167.

Biographies

Mohsen Khalily (m.khalily@surrey.ac.uk) is currently a senior lecturer (associate professor) at the University of Surrey. His research interests include surface electromagnetics, mmWave and THz technology, and antennas and propagation.

Okan Yurduseven is currently a reader (associate professor) at Queen's University Belfast. His research interests include microwave and mmWave imaging, MIMO radar, antennas, and metamaterials.

Tie Jun Cui is currently the Chief Professor at Southeast University, China. His research interests include electromagnetic fields, microwave and millimeter-wave technology, metamaterials, and metasurfaces.

Yang Hao is a professor of antennas and electromagnetics at Queen Mary University of London. His research interests include electromagnetics, metamaterials, and antennas and propagation.

George Eleftheriades is a professor at the University of Toronto. His research interests include electromagnetics, metamaterials, antennas, and metasurfaces.

CALL FOR PAPERS

IEEE COMMUNICATIONS MAGAZINE

FEATURE TOPIC: EMERGING COMMUNICATIONS TECHNOLOGIES FOR SMART CITIES AND SMART VILLAGES: DEVICES, CIRCUITS, AND SYSTEMS PERSPECTIVE

BACKGROUND

In response to the need for sustainable urban development as populations rise and cities draw ever-larger populations, smart cities are sprouting across the globe. The technologies that appear to be essential to a vision for smart cities are embedded and cyber physical systems, the Internet of everything (IoE), including Industrial Internet and Internet of Vehicles, smart energy systems, building management systems, cyber security and so on. It is anticipated that the services offered by smart city initiatives will vertically balance growth with improved and effective resource management and systems. The most recent development of ICT for Smart Cities addresses technologies, circuits, devices, systems, implementation, deployment, and applications. Like the studies/research on smart cities, the smart villages have also been attracted by the topic of how and in which way ICT can improve well-being in rural areas. By bringing these together, the aim of this Feature Topic (FT) is to present the technological advancement and challenges of smart cities and smart villages from the system design and implementation point of view. Prospective authors are invited to submit articles on topics including, but not limited to:Topics of interest include, but are not limited to:

- Intelligent Communication Systems
- Novel Communication circuits and devices (low power, secure communications, etc)
- Internet of Things (IoT) Technologies and systems
- Computing and data management-based communication networks
- M2M Communications
- Green communication technologies
- Assistive Technology for Communications
- Smart home networks and devices
- Smart healthcare devices and systems
- Smart grid systems
- Innovative design and implementation methodologies for communications
- Mission Critical communications
- Smart Communication Infrastructure in farming, fabrics. Connected health, logistics etc.
- Smart retail and banking

SUBMISSION GUIDELINES

Manuscripts should conform to the standard format as indicated in the Manuscript Submission Guidelines in the *IEEE Communications Magazine* website. Please, check these guidelines carefully before submitting since submissions not complying with them will be administratively rejected without review.

All manuscripts to be considered for publication must be submitted by the deadline through Author Portal. Select the "Series Design and Implementation of Devices, Circuits, and Systems" topic from the drop-down menu of topics and indicating "SmartCity" in Keywords. Please observe the dates specified here below noting that there will be no extension of the submission deadline.

IMPORTANT DATES

Manuscript Submission Deadline: 30 December 2022
Decision Notification: 15 May 2023
Final Manuscript Due: 31 May 2023
Publication Date: Third Quarter 2023

GUEST EDITORS

Vyasa Sai (Lead Editor)
Intel Corporation, USA
vyasa.sai@ieee.org

Mohammad Abdul Matin
North South University, Bangladesh
mohammad.matin@northsouth.edu

ACCEPTED FROM OPEN CALL

High-Data-Rate Long-Range Underwater Communications via Acoustic Reconfigurable Intelligent Surfaces

Zhi Sun, Hongzhi Guo, and Ian F. Akyildiz

The authors present a new hardware design to realize acoustic RIS, based on which the underwater RIS operation protocols are developed to address the aforementioned challenges.

Abstract

Despite decades-long development, underwater communication systems still cannot achieve high data rates and long communication ranges at the same time (i.e., beyond 1 Mb/s and 1 km). Currently, acoustic communication is the only choice to achieve long distances. However, the inherent low acoustic bandwidth results in extremely low data rates. In this article, the acoustic reconfigurable intelligent surface (RIS) system is proposed to realize high-data-rate long-range underwater communications. Although the EM-based RIS has been widely investigated in terrestrial scenarios in recent years, the underwater acoustic RIS is based on completely different physics principles. Hence, the EM-based terrestrial RISs do not work for underwater acoustic signals. Moreover, the long acoustic wave propagation delay, the inherent wideband nature, and the water behavior all impose unique challenges in underwater RIS operation. Therefore, this article presents a new hardware design to realize acoustic RIS, based on which the underwater RIS operation protocols are developed to address the aforementioned challenges. The proposed acoustic RIS system can be considered as an underwater infrastructure that enables beamforming functionalities for all types of devices, especially small robots and low-cost sensors. The simulation results show that the proposed acoustic RIS system can efficiently reflect acoustic waves and dramatically increase communication data rates and distances.

Introduction

Underwater communication is essential for future oceanic information technologies, ranging from underwater Internet of Things (IoTs) to underwater navigation and communication networks [1, 2]. Considering the unique characteristics of underwater applications and environments, there are two key requirements for future underwater communications:

1. *High data rates* for real-time and multimedia information sharing among large numbers of wireless users
2. *A long communication range* for effective coverage of the vast 3D oceanic space

State of the Art and Key Problems

Despite their important role and decades of development, existing underwater communication systems still cannot achieve *high data rates* and *long ranges* at the same time (i.e., beyond 1 Mb/s and 1 km). Specifically, first, the widely used acoustic technique can achieve several kilometers communication range. However, the data rate is limited to the kilobits-per-second level due to the narrow acoustic bandwidth and strong multipath fading [3–5]. Second, the optical communication paradigm that has attracted significant research effort in recent years is able to achieve megabits-per-second or even higher data rates. However, optical rays are not feasible to reach an underwater range of more than a few hundred meters, which limits its applicability considering the scale of the oceans/lakes. Moreover, optical signals are easily scattered by small particles in the sea and hence cannot efficiently penetrate turbid water [6, 7]. Third, electromagnetic (EM) communication paradigm utilizes medium frequency (MF) or low frequency (LF) EM signals that penetrate any type of water and provide reliable channels [1, 8]. However, due to low carrier frequencies and water absorption, the data rate and communication range of EM techniques are both limited.

To sum up, the acoustic technique is the only choice that provides a sufficient (i.e., kilometer level) underwater communication range. However, the low acoustic bandwidth causes low data rates (channel capacity), high inter-user interference, and low network throughput (network capacity). Therefore, the key question is: *Can we dramatically increase the data rates of underwater acoustic communications while maintaining its long-range advantage?*

Our Solution

To this end, in this article we propose to design an acoustic reconfigurable intelligent surface (RIS) to realize high data rates and long-range underwater communications in this article. Figure 1 shows the system architecture of an acoustic RIS-assisted underwater communication network. By using RIS-enabled beamforming with reflected signal amplification, RIS reflected paths (blue, red, and yellow solid lines in Fig. 1) provide much stronger signal-to-noise-ratio (SNR) at the

This work was supported in part by National Natural Science Foundation of China under grant no. 6227010674 and the US National Science Foundation under grant no. CNS1947748.

Digital Object Identifier: 10.1109/MCOM.002.2200058

Zhi Sun is with Tsinghua University, China; Hongzhi Guo is with Norfolk State University, USA; Ian F. Akyildiz is with Technology Innovation Institute (TII), UAE.

receiver. As a result, acoustic RIS can significantly increase the point-to-point data rate, reduce the inter-user interference, increase the network throughput, and finally, narrow the performance gap between the underwater and terrestrial wireless communications.

The proposed acoustic RIS system can be considered as an underwater information infrastructure, which enables beamforming functionality for all underwater devices in a cost-effective way. It should be noted that large-scale multiple-input multiple-output (MIMO) can also utilize the spatial resource in the same manner as the acoustic RIS does in Fig. 1. However, not all underwater devices can equip large-scale acoustic MIMO. Small devices such as underwater IoT sensors and agile swarm robots cannot afford high cost and high power consumption. In contrast, the proposed acoustic RIS does not require any hardware or software changes to end-user devices.

Research Challenges

The EM-based RIS has been intensively explored as a promising technique for 6G terrestrial wireless networks [10]. It is also proposed that the same RIS concept can also be used in other environments, such as underwater, underground, and disaster scenarios [11]. However, the research results of terrestrial RIS do not work in underwater acoustic RIS due to the fundamental differences in underlying physics and operating environments. Specifically, to realize the acoustic RIS system, four research challenges need to be addressed first:

- **Acoustic RIS hardware design:** The terrestrial RIS uses patch antennas as reflectors to control the reflected EM signals. All the RIS reconfigurability is designed based on EM and antenna theories. However, acoustic RIS needs to reflect acoustic waves, one type of mechanical waves based on completely different physics from EM waves. Hence, new hardware design is required for acoustic RIS.
- **Acoustic RIS wideband beamforming:** Due to the scarce acoustic bandwidth, underwater communication systems use ultra-wide bandwidth that is comparable to the carrier frequency. Therefore, the beamforming strategies for terrestrial RIS do not work in acoustic RIS. The wideband beamforming for acoustic RIS requires both hardware and algorithm redesign.
- **Acoustic RIS-assisted network coverage and operation:** First, terrestrial RIS only needs to cover a flat horizontal region with a vertical height no more than tens of meters. However, underwater RIS is expected to cover a vast 3D underwater space with a depth of thousands of meters. Second, the huge and dynamic acoustic signal propagation delay requires intelligent and lightweight RIS operation protocols.
- **Impacts from sea waves and turbulence motion:** Terrestrial RIS is usually installed in a static location. In contrast, underwater RIS is subject to the motions caused by sea waves and turbulence. Consequently, it is new and challenging to mitigate the impacts from RIS motions due to the uncontrollable sea waves and turbulence.

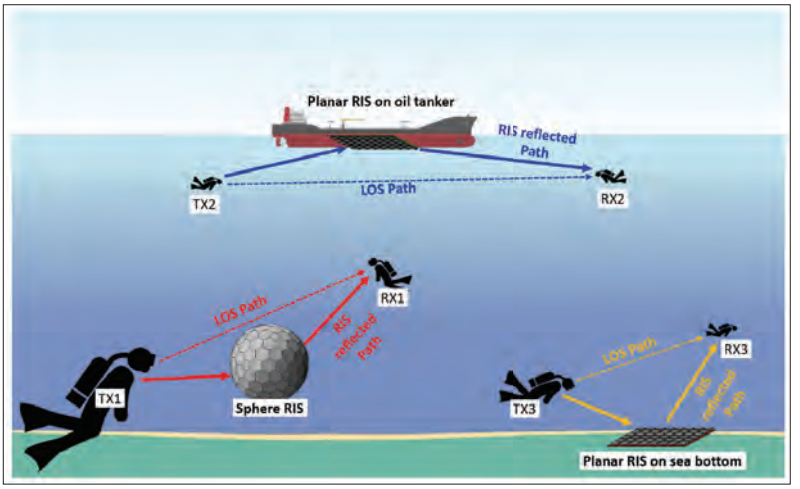


FIGURE 1. Underwater communication network using acoustic RIS as infrastructure, leading to long-range high-throughput underwater information sharing among all types of devices.

Major Contributions

In this article, we present the design and operation framework of an acoustic RIS system, aiming to achieve both high data rates and long communication ranges. All four aforementioned research challenges are fully addressed. Specifically, we first propose a hardware design to realize acoustic RIS based on an array of reconfigurable piezoelectric reflectors. Quantitative design guidelines are provided based on rigorous theoretical modeling. Second, we develop the wideband acoustic RIS beamforming scheme by jointly considering the RIS hardware and the underwater channel. Third, we propose a 3D network architecture to fully utilize the acoustic RIS to effectively cover vast under ocean space. Lightweight RIS operation protocols are also developed, which do not require any change of existing communication protocols at end users. Fourth, an extended Kalman filter (EKF)-based motion prediction algorithm is developed to mitigate the impacts from sea motion. The proposed hardware performance is validated through COMSOL Multiphysics [12], while the proposed communication and networking performance is evaluated through the Bellhop underwater simulator [13].

Acoustic RIS Hardware Design

Acoustic RIS Design using Piezoelectric Reflector Array

To develop a new RIS design paradigm for underwater acoustic signals, we propose to use the array of reconfigurable piezoelectric reflectors, as shown in Fig. 2. The illustrated RIS structure is a planar $n \times n$ array of the RIS units, as shown in the top row in Fig. 2. It should be noted that the actual RIS can have various structures, such as spherical arrays or polyhedron arrays shown in Fig. 1, depending on the application and network architecture. Also, the RIS can be placed close to the transmitter or receiver depending on the performance requirement and the environment.

Each array unit is a reconfigurable acoustic reflector, which utilizes the piezoelectric effect to effectively reflect incoming acoustic signals as well as to control the reflected acoustic signals by a microcontroller, as shown in the middle row in Fig. 2. The incident acoustic signal applies pres-

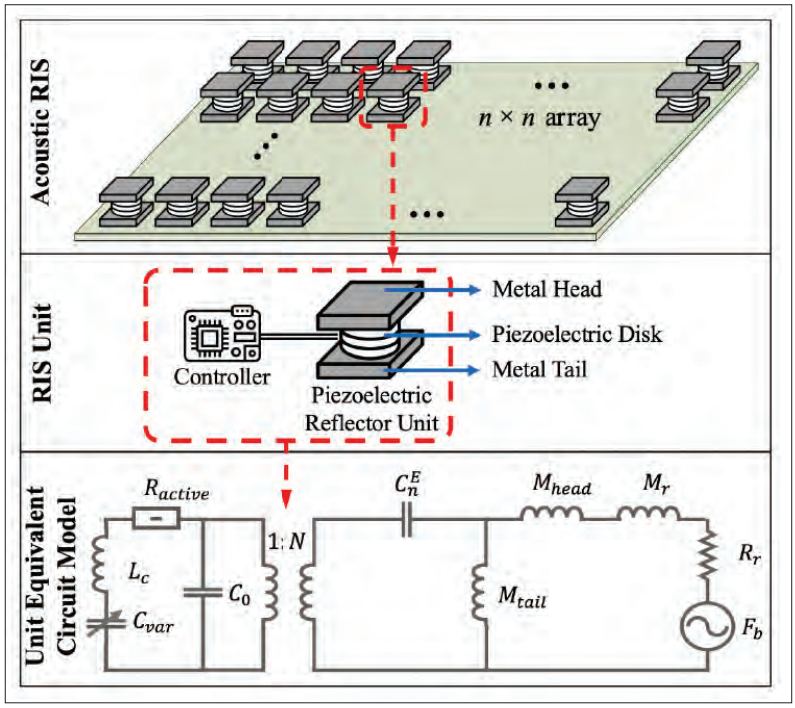


FIGURE 2. Acoustic RIS hardware design based on an array of reconfigurable piezoelectric reflectors. Each RIS reflector unit is independently controlled, and can be precisely modeled by an equivalent circuit.

sure on the metal head, which is connected with carefully designed piezoelectric rings. The core idea is using the microcontroller to change the resonant impedance of the piezoelectric reflector (a variable capacitor in our case). Thus, the elasticity of the acoustic RIS unit can be continuously tuned by the microcontroller in real time. When the elasticity of the acoustic RIS unit changes, the phase of the reflected acoustic signal changes accordingly. Specifically:

1. When the control circuit is resonant, no additional phase is imposed on the reflected acoustic signals.
2. When the circuit is off resonant, additional phase is introduced to the reflected acoustic signals.

Modeling the Reconfigurable Piezoelectric Reflector Array

Our RIS unit model is based on the equivalent circuit of a piezoelectric reflector, as shown in the bottom row in Fig. 2. The equivalent circuit is derived based on the following logic. First, the acoustic pressure of the incident sinusoidal signal applies on the metal head of one reflector unit in the RIS, which generates a force F_b. Second, the acoustic force F_b causes the whole reflector structure to move at velocity v. Third, according to Newton's Law, F_b is a linear function of v. The slope of the linear function is an equivalent mass of the system, which is determined by:

1. The mass of the metal head and tail
2. The electrical force from the piezoelectric rings
3. The interactions with the surrounded water medium

Based on the above logic model of the key factors, the equivalent circuit can be formulated as follows.

1. The acoustic force F_b is modeled as the voltage source of the circuit, while the velocity v is modeled as the current passing through the voltage source.
2. The mass of metal head M_{head} and tail M_{tail} are modeled as inductors.
3. The electrical force from the piezoelectric rings is modeled as a 1:N transformer with mechanical compliance C_n^E. The phase of the reflected signals is then controlled by the variable capacitor C_{var}.
4. The influence of the surrounded water medium is modeled as a resister R_r and an inductor M_r.

For a passive acoustic reflector unit, there is no gain applied to the reflected signals. For an active reflector unit, a negative resistance R_{active} [14] is added to the control circuit, which can give an additional gain G to the RIS reflected signals.

By using the above equivalent circuit model, the proposed reconfigurable reflector unit is fully characterized, upon which the acoustic RIS operating framework is developed.

Acoustic RIS Operation Framework

We propose an operating framework that contains four major components:

1. RIS-assisted network architecture
2. Wideband RIS beamforming design
3. Ocean movement mitigation strategy
4. Lightweight operation protocols

Underwater 3D RIS-Assisted Network Architecture

Figure 1 illustrates the RIS-assisted underwater network architecture. The acoustic RISs are deployed in the 3D underwater environment as the infrastructure to support all user devices entering this region. For each underwater communication TX-RX pair, at least one RIS is associated. By the beamforming functionality provided by the RIS, a RIS reflected path is created from TX to RIS and from RIS to RX, seen as the blue and red paths in Fig. 1. The RIS reflected path can have significant signal strength since the beamforming mechanism aligns all the reflected signals from every RIS unit. The signal after beamforming combines the contributions from all RIS units. As a result:

1. The in-band SNR is dramatically increased.
2. The effective signal bandwidth (the bandwidth within which SNR is above a threshold) is much larger.

Different from the terrestrial RIS that only needs to support 2D networks, underwater RIS is required to support user devices distributed in the vast 3D underwater space. To this end, in addition to the conventional planar RIS structure, we introduce sphere and polyhedron RIS structures, as shown in Fig. 1. The sphere and polyhedron RIS can simultaneously support horizontal and vertical communication directions in the 3D underwater space. Compared to planar RIS, the sphere and polyhedron RIS requires more complicated beamforming algorithms since the angle-of-arrival (AoA) and angle-of-departure (AoD) of adjacent RIS units could be different from each other. Fortunately, the phase for each RIS unit to form a directional beam can be accurately determined based on the geometry information of new sphere/polyhedron structures.

Wideband RIS Beamforming Design

Underwater acoustic communications use low carrier frequencies (a few to tens of kilohertz). The acoustic signal bandwidth is in the same order

of magnitude as the carrier frequency. Hence, the underwater acoustic communication is inherently wideband, which prevents the direct usage of all existing RIS beamforming solutions.

To this end, we investigate the wideband RIS beamforming strategy based on the characteristics of the proposed acoustic RIS hardware. We address two signal dispersion impacts in the wideband RIS beamforming process: the dispersion among the reflector units in an RIS and the dispersion within one reflector unit. Specifically:

- **Dispersion among reflector units:** Phase arrays realize beamforming by controlling the phase difference between adjacent array units so that the wavefronts of all reflected signals have the same direction. However, it only works in the narrow bandwidth case. For wideband signals, one set of phase assignments only guarantees a narrowband frequency form a beam toward the desired direction; other signals with different frequencies would have different propagation directions.
- **Dispersion within one reflector unit:** According to the proposed hardware design, one configuration of the reflector unit (the value of the variable capacitor C_{var} in our case) only gives the desired reflection phase to the signals within a narrow frequency band. Signals with other frequencies would have a different reflected phase from the acoustic RIS.
- **Wideband beamforming solution:** As long as the above two dispersion impacts are mitigated, acoustic wideband beamforming can be realized. Therefore, *the key idea of the proposed wideband beamforming solution is to let the above two dispersion effects cancel each other.* Since the dispersion within one reflector is the function of a piezoelectric circuit, we carefully choose the parameters in the piezoelectric circuit of each RIS unit based on the hardware model presented earlier. The objective of the circuit parameter selection is to align each single unit's signal dispersion with the dispersions across all the array units. As a result, although signal dispersions still exist, the RIS reflected signals with different frequencies have the wavefronts toward the same direction.

Ocean Movement Mitigation Strategy

Although terrestrial RIS systems are generally considered to be static, underwater acoustic RIS cannot be 100 percent static due to ocean movements such as currents, waves, and tides [6]. The ocean movements could significantly change the beam direction and reduce the data rates dramatically.

Therefore, ocean movement mitigation strategy is of great importance. As discussed in the protocol design, the acoustic RIS only needs two parameters to realize all needed RIS functionalities including the AoA of the incident acoustic signals from TX and the AoD of the reflected signals to RX. Hence, it is the relative angular movements among TX, RIS, and RX that influence the system performance. We can prove that the relative angular movement is dominated by the RIS rotation while the position changes of the three devices have negligible influence. Specifically, within the time duration of each packet transmission (a few seconds), the RIS, TX, and RX may move no more than 10 m. Considering the typical underwater transmission distance (kilometer scale), a few meters of position changes do not incur obvious relative angular movement among RIS, TX, and RX. In contrast, a small rotation of RIS can significantly change the AoA and AoD. Hence, the objective of the ocean movement mitigation strategy is to estimate the dynamic AoA and AoD in real time, and then update the RIS beamforming parameters accordingly.

In our RIS operation protocol design, the real-time AoA can be measured at most time during each duty cycle. Hence, the mitigation strategy focuses on the real-time AoD estimation. To this end, we introduce a low-cost gyroscope sensor as a part of the acoustic RIS hardware. On one hand, the sensor can measure the rotations of RIS in real time. On the other hand, the RIS can use its reflector array to measure the AoD based on the feedback signals from RX once every duty cycle. Therefore, we utilize a Kalman filter to estimate the real-time AoD using the real-time measurements from the gyro sensor and the periodically updated AoD.

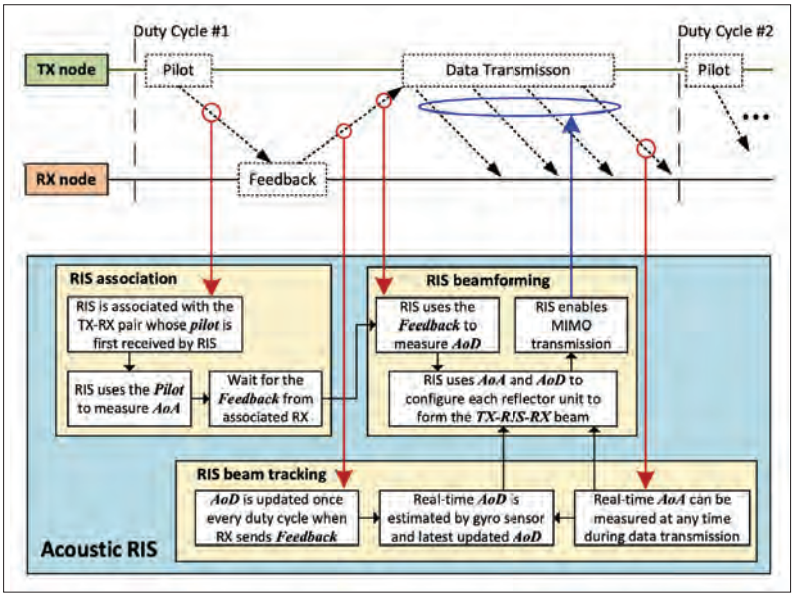


FIGURE 3. RIS operation protocol. The RIS operation (bottom blue block) is decoupled with the top underwater communication process flow.

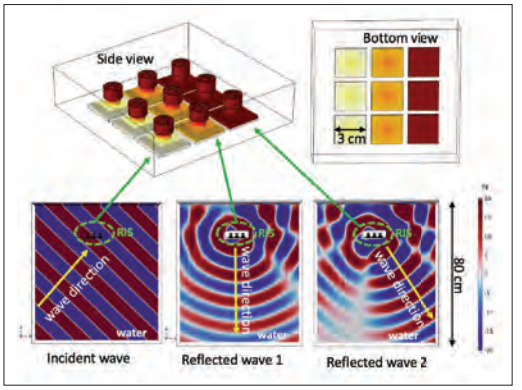


FIGURE 4. COMSOL simulation setup and results of a 3 × 3 acoustic RIS. Top: 3D illustration of the COMSOL setup. Bottom: incident and reflected sound pressure. The two different reflected sound waves are generated by adjusting the variable capacitors in the control circuits.

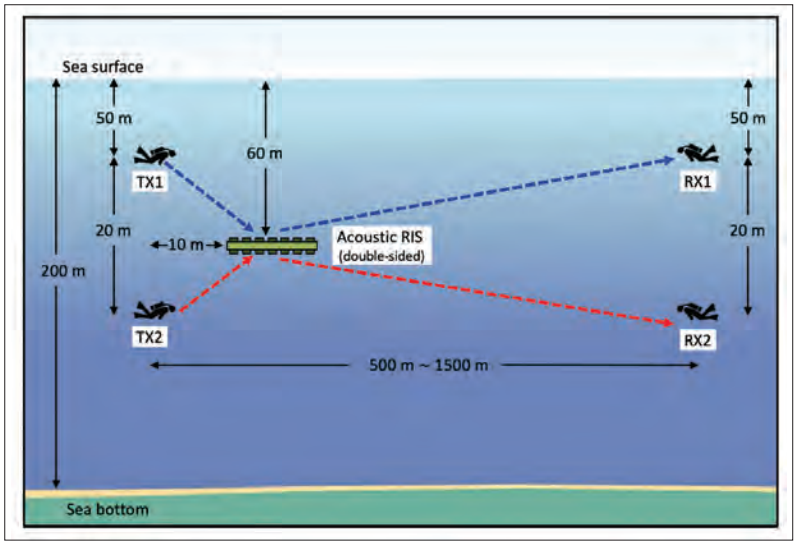


FIGURE 5. Illustration of the simulated end-to-end communication scenario. The communication consists of two pairs of TX-RX nodes. A double-sided acoustic RIS is deployed in the middle of the underwater channel.

RIS Operation Protocol

The objective of RIS operation is to enable beamforming capability for any conventional underwater single-input single-output (SISO) devices while not requiring any modification of existing underwater network protocol. Specifically, the proposed protocol decouples the RIS operations from the underwater network that the RIS supports; that is, the acoustic RIS system is an independent and decoupled infrastructure that supports the underwater networks. The underwater devices do not need to be aware of the RIS infrastructure. Hence, existing underwater communication protocols directly work without any modifications. The acoustic RIS passively overhears the controlling signals of the supported underwater network so that necessary information is collected. After that, RIS forms a beam to create a reflected path from TX node to RIS and then from RIS to RX node. During data transmissions, RIS also tracks the beam in real time so that the impacts from ocean movements are mitigated.

It should be noted that the proposed acoustic RIS is different from a full-duplex acoustic modem: the acoustic RIS directly reflects the incident waves toward the receiver without digital data processing. Each piezoelectric reflector in the RIS reflects the incident acoustic wave directly, automatically, and immediately. Moreover, the acoustic RIS can measure the phase of the incident signal while actively reflecting the signal, since the piezoelectric reflector is independent of the measurement circuit.

The RIS operation protocol is illustrated in Fig. 3, from which we can see the RIS operation (blue block on the bottom) is decoupled with the communication between TX and RX node (flow graph on the top). Without loss of generality, we consider a typical communication procedure. The RIS only overhears (red arrows) signals from TX/RX nodes, and supports the communication through beamforming (blue arrow). The RIS operation protocol consists of three main components/steps:

- **RIS association:** We assume that the multi-user access has already been determined by the supported underwater network protocol (e.g., through random access, TDMA, OFDMA) before the RIS association. An RIS continuously monitors the underwater channel and associates with the TX-RX pair whose pilot signal arrives at the RIS first. After that, the RIS focuses on the control signals and data only from that pair of TX and RX.
- **RIS beamforming:** Since the reflector array can estimate the angle of incident signal using the MUSIC algorithm [15], the RIS can derive the AoA from TX to RIS through the pilot signal from TX, and derive the AoD to RX through the feedback signal from RX. Once AoA and AoD are available, the RIS can perform beamforming to create a strong reflected path from TX to RIS and from RIS to RX, which enables the amplified beamforming functionality for any types of underwater devices. Hence, the RIS operation is lightweight and robust.
- **RIS beam tracking:** To mitigate the impacts of waves and currents, the direction of the formed beam needs to be able to track the relative angular movements among TX, RIS, and RX. Since AoA can be updated at any time during data transmission, AoA is always fresh and accurate. However, AoD can only be measured once per duty cycle when feedback signal is received, as shown in Fig. 3. Hence, we use the Gyro sensor and Kalman filter introduced earlier to estimate the real-time AoD. Then the estimation results are used to update the beamforming parameters.

System Performance Analysis

Hybrid Underwater RIS Simulator Setup

COMSOL Multiphysics is a finite element simulator that can simultaneously characterize multiple physics processes that are coupled together. We use COMSOL Multiphysics to precisely simulate:

1. The process of the reflection of acoustic signals on the RIS reflector array
2. The corresponding interactions between the mechanical displace of the piezoelectric rings and the electric current through the piezoelectric circuit

However, finite element simulators bring high computation burden. It is not feasible to use COMSOL to simulate the end-to-end underwater communications where acoustic signals propagate for thousands of meters. Thus, except in the vicinity of acoustic RIS, we use the Bellhop ray tracer to simulate the underwater channel from TX to RIS and from RIS to RX.

Acoustic RIS Simulations

Figure 4 shows the COMSOL setup and simulation results of a 3×3 acoustic RIS, aiming to validate the effectiveness of the proposed acoustic RIS. The simulation uses the AC/DC and Acoustic modules in COMSOL Multiphysics. We modified the Tonpilz Transducer Array for Sonar Systems, which was designed at 10 kHz [12]. According to the hardware design given earlier, each reflector unit has an aluminum head, piezoelectric rings, and a steel tail. As shown in the top row of Fig. 4, the aluminum head has a square shape with 3 cm edge. The interval between centers of adjacent reflector units is set to 3.5 cm.

The simulation results are given in the bottom row in Fig. 4, which are the simulated pressure of

reflected acoustic waves on the x–z plane. We apply an incident plane wave with direction (x = 1, y = 0, z = 1) and a magnitude of 100 Pa. The RIS is placed about 20 cm below the water surface with the aluminum heads facing down. We apply two sets of capacitor values to the reflector units' control circuits, aiming to form reflected beams toward two different directions, which is validated by the simulation results. Specifically, we can clearly observe that reflected wave 1 and reflected wave 2 propagate toward two different directions. It should be noted that in Fig. 4 the reflected wave is weaker than the incident wave. It is due to the small size (3 × 3) of the simulated acoustic RIS. We can see that the reflected signals can be much stronger when larger acoustic RIS is introduced in the end-to-end communication simulations (Fig. 6).

End-to-End RIS-assisted Communication Simulations

Figure 5 illustrates the simulated end-to-end communication scenario. We consider a network with two pairs of TX-RX nodes whose geometry relationships are given in the figure. We assume the TX-RX nodes that the RIS supports have already adopted the Doppler mitigation strategies. The multipath fading is also taken care of by orthogonal frequency-division multiplexing (OFDM).

A double-sided planar RIS structure is selected, which well suits the simulated network topology. The RIS reflection coefficients are derived by COMSOL model. We consider the active RIS configuration where additional power is injected through the active element (R_{active} in Fig. 2) to amplify the reflected signals. The active input for each RIS reflector is 2.5 μW. Hence, the total power consumption is 10 mW, 40 mW, and 90 mW for the 20 × 20, 40 × 40, and 60 × 60 acoustic RIS, respectively. The other physical layer parameters used in the simulations include\ 40 kHz central frequency, OFDM modulation scheme with up to 2048 subcarriers and 256-QAM subchannel modulation, TX source level of 155 dB re μPa at 1 m, background spectrum noise level of 16.99 dB re μPa per Hz, and maximum allowed bit error rate of 10^{-3}.

First, we consider a single user (point-to-point, p2p) scenario when only TX1 and RX1 in Fig. 5 are active. In this case, there is no multi-user interference, and only the facing-up RIS takes effect. Figure 6 shows the p2p communication data rates as a function of transmission distance with and without acoustic RIS, as well as with different RIS sizes. Without RIS, the data rate between the two underwater devices is no more than 2 kb/s. In contrast, by introducing the proposed acoustic RIS and the corresponding operation protocol, the data rate can reach hundreds of kilobits per second (two orders of magnitude increase) for underwater communications with a range of a few kilometers.

Second, we activate the second communication pair in Fig. 5 (i.e., TX2 and RX2). Instead of the p2p data rate, we use the network throughput (sum of the two data rates) as the evaluation metric. As the throughput also depends on the multi-user channel access protocol, we analyze two extreme cases:

1. TX1 and TX2 are perfectly coordinated (e.g., through TDMA, FDMA, and OFDMA).
2. TX1 and TX2 transmit simultaneously without any coordination.

In the first case, since all the communication pairs perfectly share the channel, the network throughput is the same as the p2p data rates given in Fig. 6. However, in the second case, due to the inter-user interference, the communication without RIS is completely ceased, resulting in zero throughput in the no-RIS scenario. In contrast, even though inter-user interference also has negative impacts, the underwater networks using RIS can still achieve a throughput of tens of kilobits per second.

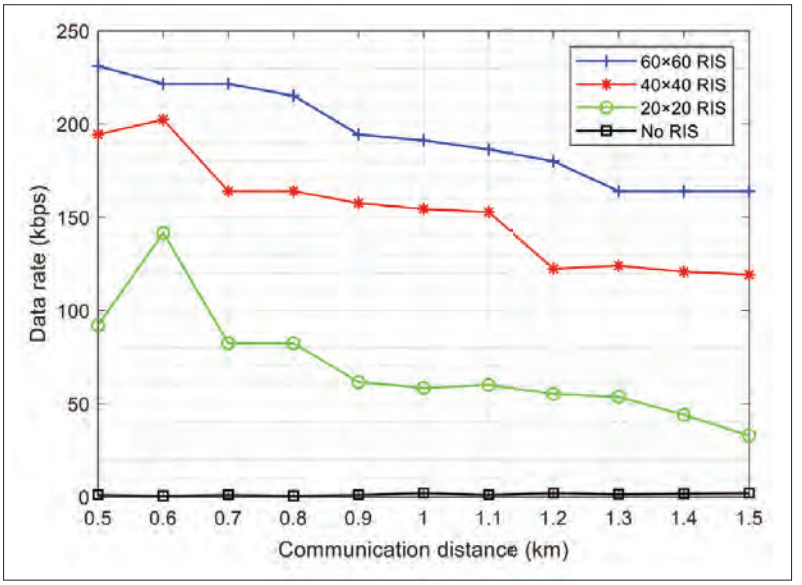


FIGURE 6. Data rates of the simulated point-to-point underwater communications between TX1 and RX1 with/without acoustic RIS, as well as with different RIS sizes.

Conclusion and Future Research

In this article, we introduce a new underwater communication infrastructure, the underwater acoustic RIS. Different from the terrestrial EM-based RIS, the underwater acoustic RIS requires completely new hardware and encounters unique challenges, such as wideband RIS beamforming, ocean movement impacts, and long signal propagation delay. We address the research challenges by designing brand-new acoustic RIS hardware and operation protocol. Through both theoretical analysis and COMSOL-Bellhop simulations, we prove that the proposed acoustic RIS system can simultaneously achieve both high data rates and a long communication range in underwater networks for any types of devices.

It should be noted that although we prove the feasibility of using acoustic RIS to significantly increase the underwater channel capacity, the actual achievable data rate also depends on the specific underwater environment and how effectively the receiver can utilize the provided received signal power. Therefore, the future research directions range from efficient RIS hardware optimization to robust RIS protocol design. First, to keep the RIS power consumption low while giving sufficient SNR at the receiver, the RIS hardware needs to be optimized to maximize the reflection efficiency. Second, multipath fading and Doppler effects are major research challenges for underwater communications. While this article is built upon state-of-the-art underwater fading and Doppler mitigation strategies, the unique new

challenges and solutions to address such problem from the RIS aspect is worth deep investigation. Third, a stable and finely calibrated deployment of an RIS system can achieve optimal performance but requires high-cost hardware and labor. Thus, the trade-off between deployment cost and performance need to be analytically modeled and optimized for different underwater applications. Fourth, new underwater receiver signal processing algorithms are needed to effectively utilize the high SNR signal enabled by the acoustic RIS.

References

[1] I. F. Akyildiz, P. Wang, and Z. Sun, "Realizing Underwater Communication through Magnetic Induction," *IEEE Commun. Mag.*, vol. 53, no. 11, Nov. 2015, pp. 42–48.

[2] A. Song, M. Stojanovic, and M. Chitre, "Editorial Underwater Acoustic Communications: Where We Stand and What Is Next?," *IEEE J. Oceanic Engineering*, vol. 44, no. 1, 2019.

[3] R. Ghaffarivardavagh *et al.*, "Ultrawideband Underwater Backscatter via Piezoelectric Metamaterials," *Proc. Annual Conf. ACM SIG on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication*, 2020, pp. 722–34.

[4] M. Stojanovic and J. Preisig, "Underwater Acoustic Communication Channels: Propagation Models and Statistical Characterization," *IEEE Commun. Mag.*, vol. 47, no. 1, Jan. 2009, pp. 84–89.

[5] J. Kusuma *et al.*, "Pragmatic Performance Optimization of a Multichannel DFE System for a Wideband 100 kb/s 1-km Subsea Acoustic Modem," *2018 4th Underwater Commun. and Networking Conf.*, 2018.

[6] C. J. Carver *et al.*, "Amphilight: Direct Air-Water Communication with Laser Light," *17th USENIX Symp. Networked Systems Design and Implementation*, 2020, pp. 373–88.

[7] W. Liu, Z. Xu, and L. Yang, "Simo Detection Schemes for Underwater Optical Wireless Communication Under Turbulence," *Photonics Research*, vol. 3, no. 3, 2015, pp. 48–53.

[8] Y. Li *et al.*, "A Survey of Underwater Magnetic Induction Communications: Fundamental Issues, Recent Advances, and Challenges," *IEEE Commun. Surveys & Tutorials*, vol. 21, no. 3, 2019, pp. 2466–87.

[9] C. Liaskos *et al.*, "A New Wireless Communication Paradigm through Software-Controlled Metasurfaces," *IEEE Commun. Mag.*, vol. 56, no. 9, Sept. 2018, pp. 162–69.

[10] C. Pan *et al.*, "Reconfigurable Intelligent Surfaces for 6G Systems: Principles, Applications, and Research Directions," *IEEE Commun. Mag.*, vol. 59, no. 6, June 2021, pp. 14–20.

[11] S. Kisseleff, S. Chatzinotas, and B. Ottersten, "Reconfigurable Intelligent Surfaces in Challenging Environments: Underwater, Underground, Industrial and Disaster," *IEEE Access*, vol. 9, 2021, pp. 150,214–33.

[12] COMSOL Inc., "Comsol," 2021; https://www.comsol.com/model/tonpilz-transducer-array-for-sonarsystems-55891.

[13] P. Qarabaqi and M. Stojanovic, "Statistical Characterization and Computationally Efficient Modeling of a Class of Underwater Acoustic Communication Channels," *IEEE J. Oceanic Engineering*, vol. 38, no. 4, 2013, pp. 701–17.

[14] E. Ugarte-Munoz *et al.*, "Stability of Non-Foster Reactive Elements for Use in Active Metamaterials and Antennas," *IEEE Trans. Antennas and Propagation*, vol. 60, no. 7, 2012, pp. 3490–94.

[15] R. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Trans. Antennas and Propagation*, vol. 34, no. 3, 1986, pp. 276–80.

Biographies

Zhi Sun [SM] (zhisun@ieee.org) received his Ph.D. degree from Georgia Institute of Technology in 2011. Currently he is a tenured associate professor at Tsinghua University, Beijing, China, which he joined in 2021. Prior to that, he was a tenured associate professor at the University of Buffalo, State University of New York, which he joined in 2012 as an assistant professor. He received the U.S. NSF CAREER Award in 2017. He is an Editor of *IEEE Transactions on Wireless Communications* and *Computer Networks* (Elsevier). His research interests lie in underground and underwater wireless communications and networking, as well as physical layer security.

Hongzhi Guo [M] (hguo@nsu.edu) received his Ph.D. degree from the University of Buffalo in 2017. Currently, he is an assistant professor at Norfolk State University. His broad research agenda is to develop the foundations for wireless sensor networks and networked robotics to automate dangerous, dirty, dull tasks in extreme environments. He received the NSF CRII award, the Jeffress Trust Awards Program in Interdisciplinary Research, and the NSF HBCU-UP RIA award.

Ian F. Akyildiz [LF] (ian.akyildiz@tii.ae) received his B.S., M.S., and Ph.D. degrees in electrical and computer engineering from the University of Erlangen-Nürnberg, Germany, in 1978, 1981, and 1984, respectively. Currently, he serves on the Advisory Board of the Technology Innovation Institute (TII) in Abu Dhabi, UAE. He is the Ken Byers Chair Professor Emeritus, and Past Chair of the Telecom group in the ECE Department (1985–2020) at Georgia Institute of Technology. He has had many international affiliations during his career. He is the Founder and Editor-in-Chief of the *ITU Journal on Future and Evolving Technologies*, the Editor-in-Chief Emeritus of Computer Networks Journal (Elsevier) (1999–2019), and was the founding Editor-in-Chief Emeritus of the *Ad Hoc Networks Journal* (Elsevier) (2003–2019). He is an ACM Fellow (1997), and has received numerous awards from IEEE, ACM, and other professional organizations.

ComSoc Technical and Emerging Technologies Committees

Where Interest, Collaboration, and Influence Connect

Join a global network of communication technology professionals advancing 33 disciplines and driving the future of technology!

Technical Committees

- Big Data
- Cognitive Networks
- Communication Theory
- Communications & Information Security
- Communications Quality and Reliability
- Communications Software
- Communications Switching & Routing
- Communications Systems Integration & Modeling
- Data Storage
- e-Health
- Green Communications & Computing
- Information Infrastructure & Networking
- InternetInternet of Things, Ad Hoc & Sensor Networks
- Molecular, Biological and Multi-Scale Communications
- Multimedia Communications
- Network Operations & Management
- Optical Networking
- Power Line Communications
- Radio Communications
- Satellite & Space Communications
- Signal Processing and Computing for Communications
- Smart Grid Communications
- Social NetworksTactile Internet
- Transmission, Access, & Optical Systems
- Wireless Communications

Emerging Technologies Initiative Committees

- Aerial Communications
- Backhaul/Fronthaul Networking & Communications
- Integrated Sensing and Communication
- Machine Learning for Communications
- Network Intelligence
- Next Generation Multiple Access
- Quantum Communications & Information Technology
- Reconfigurable Intelligent Surfaces

Volunteer Today:

www.comsoc.org/about/committees/technical-committees

IEEE ComSoc® IEEE Communications Society | IEEE

ADVERTISERS INDEX

CALL FOR SERIES TOPIC PAPERS

IEEE COMMUNICATIONS MAGAZINE

Artificial Intelligence and Data Science for Communications

Provides a forum across industry and academia to advance the development of network and system solutions using data science and artificial intelligence.

Design and Implementation of Devices, Circuits, and Systems

Provides insights into the latest technological developments in devices, circuits and systems that are used in communication systems/applications of all types, from home/consumer networks to service-provider networks.

Internet of Things

Explores the concepts of IoT and sensor networks, highlights the recent activities and achievements therein, as well as provides insights into the theoretical and practical matters related to breakthroughs in this field from different perspectives.

Military Communications and Networks

Brings together the most recent advances in military communications and networks. Papers focusing on research, development, experimentation, and deployment are solicited.

Mobile Communications and Networks

Selects and publishes in-depth, cutting-edge tutorial articles on state-of-the-art technologies and solutions for mobile wireless systems and networks, emphasizing novel but practical solutions and emerging topics of interest to industry, academia and government.

Network Softwarization and Management

Publishes articles on the latest developments in this well-established and thriving discipline, providing a forum for the publication of both academic and industrial activities, addressing the state of the art, theory and practice.

Optical Communications and Networks

Publishes top-quality, high-impact, original and unpublished articles in all areas of optical communications and networking, covering research, development, applications, and all other aspects of this field.

https://www.comsoc.org/publications/magazines/ieee-communications-magazine/cfp

IEEE ComSoc Publications

IEEE Communications Society (ComSoc) publications deliver timely, in-depth, technical information on a wide array of communications technology topics that directly impact business and advance research for the benefit of humanity.

Magazines

ComSoc's award-winning, peer-reviewed magazines cover the latest issues and advances in key areas such as wireless communications, standards, and global internetworking.

Journals

ComSoc's journals earn the highest marks by the Journal Citation Reports® (JCR) and include high-quality manuscripts covering state-of-the-art research in a variety of wireless and telecommunications topics.

More Offerings:

Best Readings | CTN | GCN | Pubs Digest | TCN | Tech Focus | White Papers

---

www.comsoc.org/publications

IEEE ComSoc®