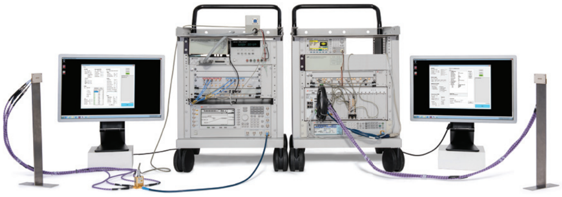
IEEE
Communications
Magazine

June 2016, Vol. 54, No. 6

- LTE-Advanced Pro
- Radio Communications
- Automotive Networking
- Bio-Inspired Cyber Security
- Consumer Communications

IEEE
IEEE ComSoc
IEEE Communications Society
A Publication of the IEEE Communications Society
www.comsoc.org



it will work here.

IEEE COMMUNICATIONS MAGAZINE

JUNE 2016, vol. 54, no. 6
www.comsoc.org/commag

LTE-ADVANCED PRO: PART 2

GUEST EDITORS: ROBERT W. HEATH JR., MICHAEL HONIG, SATOSHI NAGATA, STEFAN PARKVALL, AND ANTHONY C. K. SOONG

BIO-INSPIRED CYBER SECURITY FOR COMMUNICATIONS AND NETWORKING

GUEST EDITORS: WOJCIECH MAZURCZYK, SEAN MOORE, ERRIN W. FULP, HIROSHI WADA, AND KENJI LEIBNITZ

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: $27 per year print subscription. $16 per year digital subscription. Non-member print subscription: $400. Single copy price is $25.

EDITORIAL CORRESPONDENCE: Address to: Editor-in-Chief, Osman S. Gebizlioglu, Huawei Technologies, 400 Crossing Blvd., 2nd Floor, Bridgewater, NJ 08807, USA; tel: +1 (908) 541-3591, e-mail: Osman.Gebizlioglu@huawei.com.

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*, 3 Park Avenue, 17th Floor, New York, NY 10016.

SUBMISSIONS: The magazine welcomes tutorial or survey articles that span the breadth of communications. Submissions will normally be approximately 4500 words, with few mathematical formulas, accompanied by up to six figures and/or tables, with up to 10 carefully selected references. Electronic submissions are preferred, and should be sumitted through Manuscript Central: http://mc.manuscriptcentral.com/commag-ieee. Submission instructions can be found at the following: http://www.comsoc.org/commag/paper-submission-guidelines. For further information contact Zoran Zvonar, Associate Editor-in-Chief (zoran.zvonar@mediatek.com). All submissions will be peer reviewed.

Consumer Communications and Networking

Series Editors: Ali C. Begen, Mario Kolberg, and Madjid Merabti

Automotive Networking and Applications

Series Editors: Wai Chen, Luca Delgrossi, Timo Kosch, and Tadao Saito

Radio Communications: Components, Systems, and Networks

Series Editors: Amitabh Mishra and Tom Alexander

Accepted from Open Call

CURRENTLY SCHEDULED TOPICS

Topic	Issue Date	Manuscript Due Date
Internet of Things (IoT)	December 2016	June 15, 2016
People-Centric Internet-of-Things	February 2017	June 30, 2016
Sustainable Incentive Mechanisms for Mobile Crowdsensing	March 2017	July 15, 2016
Fog Computing and Networking	April 2017	September 1, 2016
Network Slicing in 5G Systems	May 2017	September 15, 2016

www.comsoc.org/commag/call-for-papers

Higher. Wider. Faster.
Test solutions for 5G.

The next major step beyond LTE/LTE-Advanced (4G) sets challenging requirements. Rohde & Schwarz is a world leader in all areas of RF and microwave test and measurement equipment. As a technology expert, we have been actively involved in mobile communications since the first generation. We are committed to supporting the wireless communications industry with the solutions needed to investigate, develop and standardize 5G.

Check out our test solutions at www.rohde-schwarz.com/ad/5G

ROHDE & SCHWARZ

The Emerging Era of Fog Computing and Networking

Harvey Freeman

Tao Zhang

Over the past decade, moving computing, control, and data storage into the Cloud has been the trend. However, today Cloud computing is encountering growing challenges in meeting many new requirements in the emerging Internet of Things (IoT). Such challenges include:

Latency Constraints: The stringent latency and delay requirements of many IoT systems fall far outside what mainstream Cloud services can support. For example, industrial control systems often demand end-to-end latencies to be within a few milliseconds. Many connected vehicle, virtual reality, gaming, and real-time financial trading applications may require latencies to stay below a few tens of milliseconds.

Network Bandwidth Constraints: The vast and rapidly growing number of connected things is creating data at an exponential rate. Sending all data to the Cloud will demand prohibitively high network bandwidth. This is often unnecessary. Sometimes, it is prohibited due to regulations and data privacy concerns.

Resource-Constrained Devices: IoT will support a vast number and variety of resource-constrained devices. Many such devices will not be able to rely solely on their own resources to fulfill all their computing needs. Requiring all of them to rely on Cloud services will be unrealistic and cost-prohibitive as well, because interacting with the Cloud often requires heavy processing and complex protocols. For example, the multitude of microcomputers on a car need firmware updates, but requiring each of these resource-constrained devices to perform the heavy cryptographic processing and run the complex procedures and protocols required for direct contact with the Cloud can be cost-prohibitive and also result in a system that is excessively difficult to manage.

Uninterrupted Services without Internet Access: Many IoT devices and systems, such as vehicles, drones, and oil rigs, may have intermittent network connectivity to the Cloud, but will require non-interrupted services.

New Security Challenges in IoT: Cloud and host computing alone have difficulty meeting many new security challenges in IoT. Such challenges include, for example, keeping the security credentials and software on the vast number and variety of resource-constrained devices up to date, authenticating and protecting these devices from security attacks, and assessing the security status of large distributed systems in a trustworthy manner.

Filling these and many additional gaps in today's computing models will require a new computing and networking paradigm. This new paradigm is Fog, which distributes computing, control, storage, and networking services closer to end users. Fog is a natural extension of the Cloud, bridging the Cloud and the endpoints to make computing possible anywhere along the continuum from the Cloud down to the end users. A Fog computing platform will allow the same application to run anywhere, reducing the need for specialized applications dedicated just for the Cloud or just for the edge devices. It will enable applications from different suppliers to run on the same physical platform without mutual interference. It will provide a common lifecycle management framework for all applications, offering capabilities for composing, configuring, dispatching, activating and deactivating, adding and removing, and updating applications. It will further provide a secure execution environment for Fog services and applications. This emerging Fog computing and networking era will represent a fundamental advancement in the state-of-the-art of computing and networking.

Fog provides effective ways to overcome many limitations of the existing Cloud and host computing models. Table 1 shows, for illustration, how Fog can help address the challenges we discussed at the beginning of this column.

Fog will also enable new and potentially highly disruptive business models for computing and networking. With Fog computing and networking, routers, switches, and application servers will converge into Fog nodes. Such a transformation can significantly reshape the landscape of the networking, server, and software industries. Fog-as-a-service will enable new models to deliver services to customers. Unlike Clouds that are mostly operated by large companies who can afford to build and operate huge data centers, Fog-as-a-service will enable companies, big and small, to deliver computing, storage, and control services at different scales to meet the needs of a wide variety of customers.

Proof-of-Concept (POC) trials are demonstrating the business value and technology necessity of Fog computing. For example, Cisco recently conducted a successful POC in Barcelona, where Fog computing made smart city applications more cost-effective and manageable. Barcelona envisions deploying thousands of roadside cabinets throughout the city to optimize traffic management, energy management, and water and waste management. Before they could turn this vision into reality, the city faced two major challenges. First, the traditional approach of adding new applications by adding dedicated new gateways and servers in every roadside cabinet is no longer feasible due to limited cabinet space. Second, the siloed applications have

IoT challenges	How Fog can help
Latency constraints	Store and process data, carry out control and other time-sensitive tasks near end users.
Network bandwidth constraints	Enable hierarchical data processing along the endpoint-to-Cloud continuum, hence reducing the amount of data that needs to be sent to the Cloud.
Resource-constrained devices	Perform resource-intensive tasks on behalf of resource-constrained devices when such tasks cannot be moved to the Cloud due to any reason.
Uninterrupted service with intermittent Internet access	A local Fog system can function autonomously to ensure non-interrupted services even with intermittent network connectivity to the Cloud.
New security challenges in IoT	Provide services to, for example: 1) manage and update security credentials and software on resource-constrained devices; and 2) protect devices that cannot adequately protect themselves.

Table 1. Fog provides capabilities to address IoT challenges.

been using siloed application management systems, which made the system excessively expensive to deploy and operate. Fog computing provided a solution. A single Fog node provided a common platform at each cabinet for all services and allowed applications from different suppliers to coexist without mutual interference. It provided a unified platform to support networking, security, and lifecycle management for all applications, reducing the system costs and allowing application providers to focus on developing applications rather than dedicated hardware and software for hosting and managing their applications.

On the journey to realize the full promise of Fog computing and networking, we will encounter many new challenges. For instance:

- What will Fog architectures look like?
- What new networking capabilities will Fog enable?
- How should the Fog interact with the Cloud?
- How to support the development and lifecycle management of Fog networks, services and applications?
- How to enable scalable and manageable Fog systems and networks?
- How to secure Fog computing and networking?
- How to enable users to control their Fog services?

Addressing these challenges necessitates rethinking of the end-to-end computing and networking paradigm, and will provide a fertile ground for innovation and disruption.

To that end, major industry movers and leading academic institutions joined forces to found a global Open Fog Consortium (OpenFog) in November 2015. The objective is to develop an open Fog reference architecture and to accelerate market adoption of Fog solutions. Championed by ComSoc, IEEE has entered into a strategic affiliation with OpenFog. We will co-create and co-promote Fog computing and networking concepts and architectures. We plan to jointly sponsor an annual Fog industry event. We will also co-sponsor events, journals and their special issues on Fog. Furthermore, we will jointly identify needs for new standards required to enable Fog computing and networking, and be ready to take leadership in developing these crucial standards.

At this historic moment, as we witness the emergence of the Fog computing and networking era, please join our efforts to enable and shape this new trend. The work and fun have just begun.

BIOGRAPHY

DR. TAO ZHANG, an IEEE Fellow, is a distinguished engineer/senior director of Cisco's Corporate Strategic Innovation Group. He joined Cisco in 2012 as the chief scientist for smart connected vehicles. Since then, he has also been leading initiatives to create strategies, architectures, technology, and eco-systems for the Internet of Things (IoT) and Fog Computing. Prior to Cisco, he was chief scientist and director of Mobile and Vehicular Networking at Telcordia Technologies (formerly Bell Communications Research or Bellcore). For more than 25 years he has been in various technical and executive positions, directing research and product development for vehicular, mobile, and broadband networking. He is a co-founder and a Board director of the Open Fog Consortium, the CIO of the IEEE Communications Society (2016-17), and a founding Board director of the Connected Vehicle Trade Association (CVTA). He holds more than 50 U.S. patents and has co-authored two books: *Vehicle Safety Communications: Protocols, Security, and Privacy* (2012) and *IP-Based Next Generation Wireless Networks* (2004), both published by John Wiley & Sons.

ComSoc 2016 Election
Take Time to Vote

Ballots were e-mailed and/or postal mailed 27 May 2016 to all ComSoc members (excluding Student Members, Associate Members, and Affiliates) whose memberships were effective prior to 1 May 2016. You must have an e-ballot or paper ballot before you can vote.

VOTE NOW using the URL below. You will need your IEEE account user name/password to access the ballot. If you do not remember your password, you may retrieve it on the voter login page.

https://eballot4.votenet.com/IEEE

If you have questions about the IEEE ComSoc voting process or would like to request a paper ballot, please contact ieee-comsocvote@ieee.org or +1 732 562 3904.

If you do not receive a ballot by 30 June, but you feel your membership was valid before 1 May 2016, you may e-mail ieee-comsocvote@ieee.org or call +1 732 562 3904 to check your member status. (Provide your member number, full name, and address.)

Please note IEEE Policy (Section 14.1) that IEEE mailing lists should not be used for "electioneering" in connection with any office within the IEEE.

Voting for this election closes 22 July 2016 at 4:00 p.m. EDT! Please vote!

UPDATED ON THE COMMUNICATIONS SOCIETY'S WEB SITE

www.comsoc.org/conferences

2016

JULY

OECC/PS 2016 —Optoelectronics and Communications Conference/Int'l. Conference on Photonics in Switching, 3–7 July

Niigata, Japan
http://www.oecc-ps2016.org/

ICUFN 2016 — Int'l. Conference on Ubiquitous and Future Networks, 5–8 July

Vienna, Austria
http://icufn.org/main/

CITS 2016 — Int'l Conference on Computer, Information and Telecommunication Systems

6–8 July
Kunming, China
http://atc.udg.edu/CITS2016/

IEEE ICME 2016 — IEEE Int'l. Conference on Multimedia and Expo, 11–15 July

Seattle, WA
http://www.icme2016.org/

SPLITECH 2016 — Int'l. Multidisciplinary Conference on Computer and Energy Science, 13–15 July

Split, Croatia
http://splitech2016.fesb.hr/

SPECTS 2016 — Int'l. Symposium on Performance Evaluation of Computer and Telecommunication Systems, 24–27 July

Montreal, Canada
http://atc.udg.edu/SPECTS2016/

TEMU 2016 — Int'l. Conference on Telecommunications and Multimedia, 25–27 July

Heraklion, Greece
http://www.temu.gr/

IEEE/CIC ICCC — Int'l. Conference on Communications in China, 27–29 July

Chengdu, China
http://iccc2016.ieee-iccc.org/

ICCE 2016 — IEEE Int'l. Conference on Communications and Electronics, 27–29 July

Ha Long, Vietnam
http://www.icce-2016.org

AUGUST

ICCCN 2016 — Int'l. Conference on Computer Communication and Networks, 1–4 Aug.

Waikoloa, HI
http://icccn.org/icccn16/

ISMW-FRUCT 2016 — Int'l. FRUCT Conference on Intelligence, Social Media and Web, 28 Aug.–4 Sept.

St. Petersburg, Russia
http://ismw-fruct.spbu.ru/#general

SEPTEMBER

IEEE PIMRC 2016 — IEEE Int'l. Symposium on Personal, Mobile, and Indoor Radio Communications, 4–7 Sept.

Valencia, Spain
http://www.ieee-pimrc.org/

IEEE EDOC 2016 — IEEE Int'l. Enterprise Distributed Object Computing Conference, 5–9 Sept.

Vienna, Austrial
http://edoc2016.univie.ac.at/

ASMS/SPSC 2016 — Advanced Satellite Multimedia Systems Conference and the Signal Processing for Space Communications Workshop, 5–7 Sept.

Palma De Mallorca, Spain
http://www.asmsconference.org/

ITC28 2016 — Int'l. Teletraffic Congress, 12–16 Sept.

Würzburg, Germany
http://itc28.org/

IEEE HEALTHCOM 2016 — IEEE Int'l. Conference on e-Health Networking, Application & Services

Munich, Germany
http://ieeehealthcom2016.com/
call-for-submission

IEEE SARNOFF SYMPOSIUM 2016 — IEEE 37th Sarnoff Symposium 2016, 19–21 Sept.

Newark, NJ
http://sites.ieee.org/sarnoff2016/

ISWCS — Int'l. Symposium on Wireless Communication Systems, 20–23 Sept.

Poznan, Poland
http://iswcs2016.org/

ICACCI 2016 — Int'l. Conference on Advances in Computing, Communications and Informatics, 21–24 Sept.

Jaipur, India
http://icacci-conference.org/2016/home

SOFTCOM 2016 — Int'l. Conference on Software, Telecommunications and Computer Networks, 22–24 Sept.

Split, Croatia
http://marjan.fesb.hr/SoftCOM/2016/cfp.html

NETWORKS 2016 — Int'l. Network Strategy and Planning Symposium, 26–28 Sept.

Montreal, Canada
http://networks2016.etsmtl.ca

IEEE WISEE 2016 — IEEE Int'l. Conference on Wireless for Space and Extreme Environments, 26–29 Sept.

Aachen, Germany
http://www.ti.rwth-aachen.de/WiSEE2016

OCTOBER

IEEE CLOUDNET 2016 — IEEE Int'l. Conference on Cloud Networking, 3–6 Oct.

Pisa, Italy
http://cloudnet2016.ieee-cloudnet.org

ICMU 2016 — International Conference on Mobile Computing and Ubiquitous Networking, 4–6 Oct.

Kaiserslautern, Germany
http://www.icmu.org/icmu2016/

APNOMS 2016 — Asia-Pacific Network Operations and Management Symposium, 5–7 Oct.

Kanazawam, Japan
http://www.ieice.org/~icm/apnoms/2016/

ATC 2016 — Int'l. Conference on Advanced Technologies for Communications, 12–14 Oct.

Hanoi, Vietnam
http://rev-conf.org

WCSP 2016 — Int'l. Conference on Wireless Communications & Signal Processing, 13–15 Oct.

Yangzhou, China
http://ic-wcsp.org

–Communications Society portfolio events appear in bold colored print.
–Communications Society technically co-sponsored conferences appear in black italic print.
–Individuals with information about upcoming conferences, Calls for Papers, meeting announcements, and meeting reports should send this information to: IEEE Communications Society, 3 Park Avenue, 17th Floor, New York, NY 10016; e-mail: p.oneill@comsoc.org; fax: + (212) 705-8996. Items submitted for publication will be included on a space-available basis.

GCN GLOBAL COMMUNICATIONS NEWSLETTER

June 2016
ISSN 2374-1082

CONFERENCE REPORT

Barcelona Mobile World Congress 2016

By Juan Pedro Muñoz-Gea, Josemaría Malgosa-Sanahuja, and Pilar Manzanares-López, Universidad Politécnica de Cartagena, Spain

As everyone knows, the Mobile World Congress (MWC) is the world's biggest and most influential mobile event. This year the conference was held February 22–25 in Barcelona (Spain). This edition of MWC was the 10th time that it has been held in Barcelona. Until 2005 the Congress was held in Cannes (France) under the name of 3GSM World. In 2006 it was moved to Barcelona and since then it has become one of the biggest technology events. This relationship will continue until at least 2023, thanks to the agreement that was reached recently.

The event consists of three main blocks. The first block is the conferences, all related to mobile technologies; the second is the exhibition zone, where companies show their novelties; the third block consists of a dozen parallel events. The presence of up to 95,000 attendees and approximately 2,000 companies that have participated in the exhibition zone are both clear indicators of the success of MWC-16.

MWC is mainly focused on mobile devices. In past years, there have been huge announcements of devices at MWC. However, this year MWC was also focused on other issues, like vehicular technology, new mobile payment methods, new trends in ISP (Internet service providers) mobile products and services, and finally virtual reality, whose scale of presence was the biggest surprise. The most relevant companies in the cellular market showed their new mobile designs that incorporate a virtual reality (VR) headset. The image of conference attendees looking in all directions with their virtual reality headsets has become the snapshot that best summarizes MWC-16. Now it is time to see if the applications supporting this new technology actually attract consumers.

On the other hand, the new mobile payment services and related technologies presented at MWC-16 also attracted great attention. One of the most promising was the Paypal Here reader, based on NFC (near field communication) technology. This reader is able to understand any type of payment method. In the same business line, another important novelty was the mobile banking service called imaginBank, presented by La Caixa. With imaginBank, you can do what you usually do in a traditional bank, but also much more, thanks to the inherent benefits of IT technologies. In imaginBank, all the banking services are online; the users manage their own financial resources by themselves, with the help of a set mobile app and social networks. The final touch was led by MasterCard, who presented an authentication technology based on selfies called selfie-pay. This app enables consumers to validate their transactions with a simple selfie. It is clear that all these pieces must be put to work together in order to provide value added on-line banking and shopping services to the community.

The MWC-2016 was held in Barcelona, in one of the most beautiful exhibitions called Fira de Barcelona.

The image of conference attendees looking at all directions with their virtual reality headsets has become the snapshot that best summarizes MWC-16.

Regarding mobile devices, the most relevant premiere was a modular phone design that lets you attach accessories directly to it. The device can be prepared in any way the consumer likes it. Another interesting issue was the internal water circuits to keep the processor cool, keeping it from overheating. However, it remains to be seen if consumers continue to opt for high-end devices, or as recently noted in data sales, these types of high-performance mobile devices are passed over in favor of an increasingly compelling midrange in features and prices.

The next generation of mobile communications (5G) was on many MWC-16 minds. The focus on 5G has been all about connecting things to the Internet, when in fact much of the world does not even have access to the basics. Facebook Chief Executive Mark Zuckerberg pointed out that while a small section of the connected world is racing to embrace next-generation technology, the majority of people, including large swaths of Europe and the USA, are still using 2G, a technology that is 25 years old. Zuckerberg cautioned that the gap between the small wealthy majority and everyone else is only going to widen if we keep going the way we are.

Finally, another remarkable event at MWC-16 was the Mobile Premier Awards (MPA). The MPA is an organization that was born in Barcelona, and it is completely independent of MWC. However, this year they presented the award at MWC-16. It is an acknowledgement of the best app developed by startup compa-

(Continued on Newsletter page 4)

REGIONAL REPORT

Telekom Romania: A New Beginning? An Interview with Miroslav Majoros, CEO, Telekom Romania

By Nicolae Oaca, Romania

Telekom Romania, the former Romtelecom and Cosmote Romania, is facing network challenges in one of the most competitive markets in the EU. In the fixed line business, the former Romtelecom is competing with RCS&RDS, UPC Romania, and telcos with larger optical fiber networks, while in the mobile business, the main competitors are Orange and Vodafone, celcos with better national coverage of LTE networks and providing higher speed access. In September 2014 Romtelecom and Cosmote were rebranded under the Telekom logo, while the merger process started in 2013. On January 1, 2016, Miroslav Majoros, an executive with a telecom engineering background and with an MBA from the prestigious Harvard Business School and Stanford Graduate School of Business, came from Slovakia to turn arround Telekom Romania.

Miroslav Majoros

One of the main problems behind the Telekom evolution in the last few years was financing. How do you intend to raise funds to develop the business, mobile mainly, to reduce the gap?

The overall investment plans for 2016 amount to over €180 million. Most of the investments will be directed toward optical fiber networks and mobile networks, based on 3G and 4G technologies. The increase in the investments planned, by more than 35 percent compared to the previous year, was more than necessary, given that the local market is extremely competitive, and without putting money in the infrastructure it is impossible to succeed in the long term.

Telekom Mobile lags behind its main competitors from the point of view of LTE networks. What is your strategy to rapidly reduce the gap (with Orange, Vodafone)?

We will continue to develop our LTE infrastructure, as it is one of our strategic objectives. The agreement with Orange represents an immediate support for our plans, but in the medium and long term it is natural to continue developing our own infrastructure.

Telekom Mobile aquired only one 2x5 MHz bloc in the 800 MHz band, while Orange and Vodafone aquired two blocks. What are your intentions to keep pace with your competitors?

Depending on the evolution of the 4G mobile telecommunications market and on the future development of our own network, we might consider acquiring a supplementary block for this bandwidth.

Recently, Deutsche Telekom's top managers declared that the fiber network is a top priority for Telekom Romania. Why not the LTE network, having in mind that mobile business accounts for two thirds of Romanian telecommunications revenues?

It is not a fixed versus mobile business strategy. We started to operate having in mind an integrated approach, therefore both segments are equally relevant. We will focus on expanding both the fiber and LTE networks, on re-launching the portfolio of fixed-mobile services, and on improving the quality of services for customers.

Telekom–Orange wholesale and national roaming agreements respectively mean giving access to the Telekom fixed network, a strategic asset Orange never could have, and to the Orange LTE network, an asset Telekom already has. How do you comment?

This agreement is a good opportunity for us to provide improved services to our customers. Our main focus is represented by the fixed-mobile convergent packages, under the MagentaONE proposition. Extended coverage for both fixed and mobile networks will help us significantly increase the areas where we can offer our services to customers. Therefore, the national roam-

(Continued on Newsletter page 4)

CHAPTER REPORT

IEEE ComSoc Student Members Visit AT&T Data Center in Tlalnepantla, Mexico

By Jose-Ignacio Castillo-Velazquez, UACM, Mexico

On February 25, 2016, AT&T for the first time in México opened its data center during the AT&T High Tech Day, when ComSoc student members and student branch members from Universidad Autonoma de la Ciudad de México, with students from other universities, visited the center. AT&T opened some of its data centers that same day. Students had the chance to communicate using video conferencing with students in other cities where AT&T has data centers, such as San Juan, Puerto Rico; Texas, New Jersey and Florida in the U.S.; and Tlalnepantla in Mexico State, near Mexico City.

In a picture at AT&T facilities are Fabricio Astorga Martínez, Luis Carlos Revilla Melo, Daniel Javier Serrano Martinez, Fernando Trueba, Adrian Martinez, Alonso Delgado, and Violeta Perez from UACM, with students from other universities in Mexico City.

Until December 2014, the mobile market in México had the following distribution: TELCEL-AMERICA MOBILE (from Mexico) was the largest with 70.4 million users; second was MOVISTAR-TELEFONICA (from Spain) with 20.5 million users; IUSACELL-was third with 8.5 million users; and NEXTEL was fourth with 2.8 million users. In January 2015 AT&T (from the U.S.) re-entered the market in México, buying IUSACELL and NEXTEL. Now AT&T is the third largest competitor measured by number of users, 11.4 millions, but second in revenues and coverage with 90 percent, behind TELCEL, which covers 94 percent, and ahead of MOVISTAR, which covers 80 percent of Mexico. Students had the chance to visit the AT&T data center and its NOC (network operations center).

In the U.S., the purpose of AT&T High Tech Day was to get high school students excited about careers in the fields of science, technology, engineering, and mathematics (STEM). This event has occurred since 1998;

(Continued on Newsletter page 4)

CONFERENCE REPORT

VITEL 2015: 31st Workshop on Telecommunications Critical Infrastructure and ICT

By Tomi Mlinar and Marko Jagodic, Slovenian Electronic Communications Society, Slovenia

Round table with Nikolaj Simic, chairman (left) and other members (B. Tavcar, M. Mrzel-Ljubic, F. Dolenc, M. Turk, B. Ivanc, and V. Podlogar).

VITEL 2015, the 31st workshop on telecommunications, took place at the Congress Centre Brdo pri Kranju in Slovenia 11–12 May 2015. A Program Committee, chaired by Bostjan Tavcar, selected 'Critical Infrastructure and ICT' as a theme for the workshop. An event was organized by the Slovenian Electronic Communications Society, a member of the Electrotechnical Association of Slovenia, and sister society of the IEEE Communications Society. A group of 35 authors and co-authors prepared 22 papers, and more than 130 participants attended the workshop. A round table, chaired by Nikolaj Simic, president of the organizing committee, was dedicated to security threats in critical infrastructure and their consequences on state security. Members of the round table discussed facilities and services crucial for the country.

At the VITEL 2015 workshop, several lecturers pointed out a significant impact of discontinuity of ICT activities and operation on national security, the economy, and critical societal functions, including health, safety, personal security, and social welfare.

The European Union Council Directive on the identification and designation of European Critical Infrastructures (ECI) and the assessment of the need to improve their protection 114/2008/EC requires implementation of relevant legal measures from the member states. Regarding the action priorities or direct impact on other sectors of critical infrastructure in the EU, critical infrastructure was classified according to priority order, where information and communication support was listed as the second priority.

Thus, the VITEL 2015 workshop focused on ICT systems as important tools for the protection of critical infrastructure. Bostjan Tavcar, who in addition to serving as president of the workshop program committee also is the head of the Administration of the Republic of Slovenia for civil protection and disaster relief, opened the workshop with his introductory speech. Then the attendees were honored with the remarkable opening speech of academic professor Tadej Bajd, president of the Slovenian Academy of Sciences and Arts.

In the two days of the VITEL 2015 workshop lecturers from public institutions, research institutes, universities, and private companies, the following topics were addressed:

•A critical infrastructure for providing IT and telecommunications services and relevant solutions.

•Determination of cyber threats and vulnerabilities of critical infrastructure.

Academic Prof. Tadej Bajd, President of the Slovenian Academy of Aciences and Arts.

•Provision of telecommunications and information services in natural disasters and other emergency situations. An audience heard about actual experiences from infrastructure operators, e.g. telecoms, broadcasters, the electric power industry, and a National protection and rescue directorate from Croatia in major natural and other disasters. Cases included sleet (freezing rain) in Slovenia in February 2014 and floods in the Northern Balkans in May and June 2014.

•The role of the State and civil protection service in providing minimal functions of public telecommunications networks in natural and other disasters. Lecturers and the audience had a fruitful discussion on the question of how the Administration for Civil Protection and Disaster Relief of the Republic of Slovenia and local civil protection organizations could assist in providing telecommunications and information services in such accidents.

•Levels of reliability, availability, and security provided by new technologies.

•Functioning of the emergency call service (112) and critical infrastructure in major natural and other disasters.

•The choice of LTE as the telecommunications platform for professional radio communications.

•Development of radio networks (e.g. DMR) for the critical communications.

•The role of the national regulator in assuring non-disturbed provision of ICT services in the event of natural and other disasters. A lecturer from the Agency for Communication Networks and Services from the Republic of Slovenia focused on the question whether the Electronic Communications Act needs to be reworded or changed in relation to critical infrastructure.

•DNS as critical infrastructure. A lecturer from the Academic and Research Network of Slovenia showed how DNS (domain name server) has been involved in national critical infrastructure and what approach and measures should be taken for risk management of DNS at the national ccTLD Registry.

In addition to the aforementioned topics attendees heard several lectures related to the private software and hardware solutions used for critical infrastructure management.

Attendees finished the two days of very interesting and fruitful discussions with the conclusion that a proper functioning of the ICT in case of critical situations is the basis for all other industries, and should be given more value in the future.

After the great success of the 31st workshop on telecommunications, we are looking forward to the 32nd workshop, which will be held from 16–17 May, 2016 at the Brdo Congress Centre, Brdo pri Kranju, Slovenia. The title of the next workshop is 'Smart Networks of the Information Society'. You are kindly invited!

The 31st VITEL workshop participants in front of the Brdo Congress Centre, Slovenia.

TELECOM ROMANIA/*Continued from page 2*

ing agreement with Orange for access to its 4G and 4G+ networks will enable us to also win new customers who will thus benefit from extended 4G coverage and higher quality of 4G and 4G+ services, in addition to the fixed broadband and best-in-class TV services within Telekom's integrated bundles. We estimate that we will be able to launch the first commercial packages based on this agreement in May 2016.

How about repositioning Telekom in the Romanian market?

This is a process that started almost two years ago with the rebranding, when Telekom announced a new vision and strategy to further differentiate in a market mainly driven by price. The launch of the integrated fixed-mobile packages was the first step in this process, bringing customers an integrated communications proposition in the market and a new customer experience, with simple, transparent services and convenience — one invoice, one call center, one MyAccount.

We will continue to build on this strategy. We have a strong integrated fixed-mobile proposition, a complete service portfolio for B2B, and a very good TV offer. It is now time to focus more on the next steps of the process and our strategy for how to do this is very clear: by delivering an excellent customer experience and great value for the money to our customers. In the coming period we will therefore focus on expanding our networks, for both fixed and mobile technologies, as the base for innovative, interactive, and converged services, on consolidating our convergent service portfolio, and on improving the customer experience.

The essence of our strategy is to offer great value for our customers through bringing more benefits and competitive, simplified, and innovative services, along with providing a great customer experience through all touchpoints. The lowest price and the cheapest services are not the vision we share for a sustainable industry and for creating value to customers.

Is it part of the strategy to merge DT operations in Romania?

We are heading in this direction. The two companies are aligned operationally and are working in sync to offer an integrated customer experience and to ensure commercial consistency. However, a complete merger is a more subtle process that goes beyond procedures, functions, and structures, up to mentalities and organization culture. At this point there is still work to do in terms of harmonizing the two cultures. This is a process that takes time, no matter what actions you take and how much effort you put in.

Last but not least, the legal merger is a very complex process and is subject to different regulations and various approvals, and not only from the shareholders.

Could RomTelekom have an IPO this year?

It is a shareholders' matter to decide on what is the best solution for privatization and the proper timing. What I can tell you regarding the status of this process is that for the moment we are waiting for a decision from the Ministry upon the solution that the State will opt for, direct negotiation or IPO. In the meantime, we can only reiterate that Telekom Romania is part of the OTE and DT Groups, which are fully committed to their presence in Romania and the country's prospects. The recent rebranding and our continuous investments in the local market are evidence of this commitment.

BARCELONA/*Continued from page 1*

nies around the world. There were 16 apps competing for the award (http://mobilepremierawards.com/finalists-2016), which in the end was won by Jordi Llonch, CEO and founder of Sharing Academy, S.L. The application puts in students in contact with senior students, who act as teachers. The app was born out of Jordi's personal experience, and it receives approximately 6,000 visits every day.

It can be concluded that the most recent edition of Barcelona MWC did not strictly follow the tradition of focusing on mobile devices. On the contrary, it expanded to other very prominent fields of the mobile world. In this way, Barcelona MWC has definitely become the principal reference of the mobile technology.

STUDENT MEMBERS VISIT AT&T/*Continued from page 2*

this year was the first time it occurred in Mexico due to the recent AT&T acquisition. Because of confidentiality reasons, students could not take pictures inside the data center and NOC facilities. Students received snacks and flash memories from AT&T as souvenirs.

At the end of the visit, the students had the chance to talk with the AT&T data center's CTO (Chief Technology Officer). Now those students have a better idea how a real data center and NOC work, and they have also expanded their career opportunities to consider when they will graduate.

GCN GLOBAL COMMUNICATIONS NEWSLETTER

STEFANO BREGNI
Editor
Politecnico di Milano — Dept. of Electronics and Information
Piazza Leonardo da Vinci 32, 20133 MILANO MI, Italy
Tel: +39-02-2399.3503 — Fax: +39-02-2399.3413
Email: bregni@elet.polimi.it, s.bregni@ieee.org

IEEE COMMUNICATIONS SOCIETY

STEFANO BREGNI, VICE-PRESIDENT FOR MEMBER AND GLOBAL ACTIVITIES
CARLOS ANDRES LOZANO GARZON, DIRECTOR OF LA REGION
SCOTT ATKINSON, DIRECTOR OF NA REGION
ANDRZEJ JAJSZCZYK, DIRECTOR OF EMEA REGION
TAKAYA YAMAZATO, DIRECTOR OF AP REGION
CURTIS SILLER, DIRECTOR OF SISTER AND RELATED SOCIETIES

IEEE COMMUNICATIONS SOCIETY

MARKO JAGODIC (jagodic.marko@guest.arnes.si)
NICOLAE OACA (nicolae_oaca@yahoo.com)
JOSEMARIA MALGOSA SANAHUJA (josem.malgosa@upct.es)

IEEE ComSoc
IEEE Communications Society

www.comsoc.org/gcn
ISSN 2374-1082

OMBUDSMAN

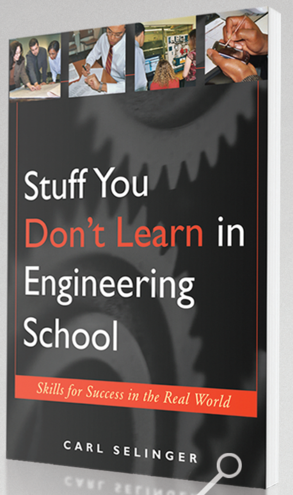
ComSoc Bylaws Article 3.8.10

"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."

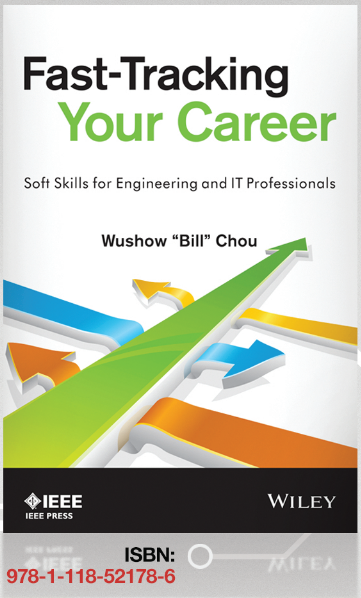
IEEE Communications Society Ombudsman
c/o Executive Director
3 Park Avenue
17 Floor
New York, NY 10017, USA
ombudsman@comsoc.org

www@comsoc.org "About Us" (bottom of page)

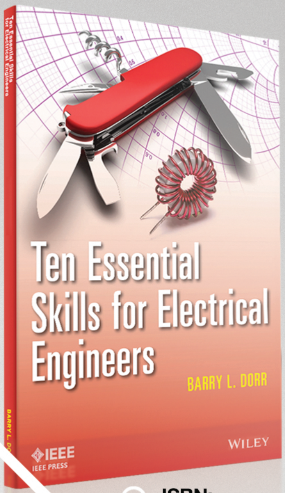
Advance Your Career
With the Experts at Wiley-IEEE Press


Stuff You Don't Learn in Engineering School
Skills for Success in the Real World
CARL SELINGER

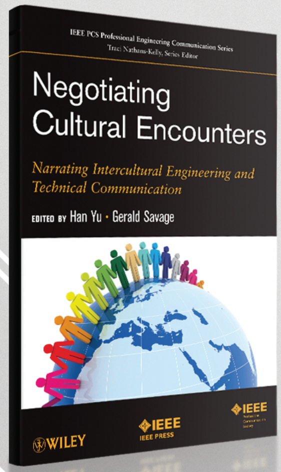
ISBN: 978-0-471-65576-3


Fast-Tracking Your Career
Soft Skills for Engineering and IT Professionals
Wushow "Bill" Chou
IEEE
IEEE PRESS
WILEY

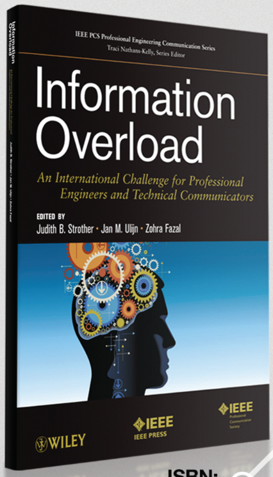
ISBN: 978-1-118-52178-6



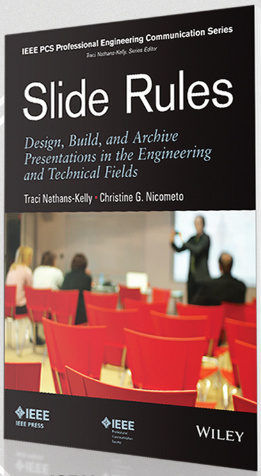
ISBN: 978-1-118-52742-9


IEEE PCS Professional Engineering Communication Series
Traci Nathans-Kelly, Series Editor
Negotiating Cultural Encounters
Narrating Intercultural Engineering and Technical Communication
EDITED BY Han Yu · Gerald Savage
WILEY
IEEE
IEEE PRESS
IEEE

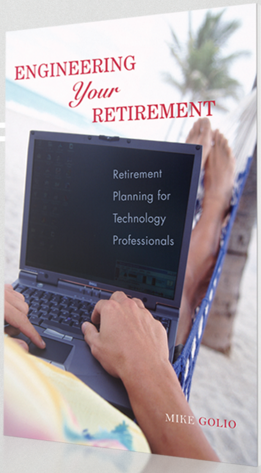
ISBN: 978-1-118-06161-9


IEEE PCS Professional Engineering Communication Series
Traci Nathans-Kelly, Series Editor
Information Overload
An International Challenge for Professional Engineers and Technical Communicators
EDITED BY Judith B. Strother · Jan M. Ulijn · Zohra Fazal
WILEY
IEEE
IEEE PRESS
IEEE

ISBN: 978-1-118-23013-8


IEEE PCS Professional Engineering Communication Series
Traci Nathans-Kelly, Series Editor
Slide Rules
Design, Build, and Archive Presentations in the Engineering and Technical Fields
Traci Nathans-Kelly · Christine G. Nicometo
IEEE
IEEE PRESS
IEEE
WILEY

ISBN: 978-1-118-00296-4


ENGINEERING Your RETIREMENT
Retirement Planning for Technology Professionals
MIKE GOLIO

ISBN: 978-0-471-77616-1

ARE YOU AN EXPERT IN YOUR FIELD?
TO LEARN MORE ABOUT PUBLISHING WITH WILEY-IEEE VISIT:

HTTP://WWW.IEEE.ORG/PUBLICATIONS_STANDARDS/PUBLICATIONS/PRESS/AUTHORS/BOOKS.HTML

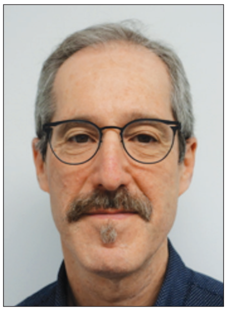
IEEE

VISIT: WWW.WILEY/IEEE.COM

WILEY

LTE-ADVANCED PRO: PART 2

Robert W. Heath Jr.


Michael Honig

Satoshi Nagata

Stefan Parkvall

Anthony C. K. Soong

Our society is undergoing an unprecedented transformation as more and more devices are being interconnected over the wide area network. This will profoundly change our productivity and human interaction. The Internet of Things (IoT) has been one of the most successful growth segments in cellular-based applications in recent years, with an annual growth rate in the range of 30 percent. It is anticipated that the ratio of connected things to people will rise sharply over the next 5–10 years, to around 7:1 or even higher, which means that there would be 50 billion or even more connected things. The GSM Association believes the number could grow to 24 billion by 2020 [1], while Gartner forecasts that number to be 35 billion [2]. Applications include smart building, smart metering, smart city, ehealth, smart environment, consumer electronics, and telematics/vehicle to everything (V2X). Consequently, this part of the Feature Topic will focus on technologies for Long Term Evolution (LTE) that will enhance the support of IoT.

Mobile networks can embrace machine-to-machine (M2M) communication services with wide-area coverage, low cost, low latency, and massive number of connections. LTE-Advanced Pro can surpass the limits of legacy fourth generation (4G) networks, which were not specifically designed for the particular requirements of some M2M scenarios (e.g., very long terminal battery life of up to several years). Such benefits make it possible for operators to generate revenue from vertical markets, especially in smart metering, vehicle communication, and wearable devices.

Industry applications such as smart metering require good coverage, up to 20 dB more than current LTE networks, and low terminal-side power consumption, up to 10 years of battery lifetime. The existing cellular network is mainly designed for capacity enhancement and is to a lesser extent focused on reaching a massive scale of connectivity, which motivates further evolution.

There are three candidate technologies, including eMTC, EC-GSM-IoT, and NB-IoT, which are being discussed and standardized in Third Generation Partnership Project (3GPP) Release 13. The first article of this part of the Feature Topic gives an overview of Release 13 for M2M communication. It elucidates physical layer as well as medium access control (MAC) and higher layer signaling features.

The second article specifically addresses the LTE evolution for V2X services. The article presents the use cases and addresses the main challenges of high mobility and densely populated vehicle environments. The main goal of this evolution is to enable the vehicles to communicate with other vehicles, pedestrians, and infrastructure in order to exchange messages for aiding in road safety (e.g., collision avoidance messages), controlling traffic flow, and providing various traffic notifications. It will no doubt alter the way that we interact with our automobiles in the future.

This third article discusses enhancements needed in LTE-Advanced Pro to enable a scalable cloud radio access network (C-RAN). The article focuses on collaborative signal processing, resource management, and a green architecture for C-RAN systems. It discusses how to take advantage of the near sparsity of the channel matrix to significantly reduce the channel estimation overhead and computational complexity as well as the BBU management, the RRH on/off problem, and problems caused by finite capacity of the transport network.

The fourth article in this Feature Topic discusses the wireless factory automation use case for LTE. The article classifies these requirements and identifies the opportunities that the current LTE air interface has for factory automation applications. It goes on to discuss the features that will be needed in its evolution to support this application for increasing factory productivity.

The fifth article gives an overview of the technical features of LTE-Advanced Pro in Release 14. It discusses the support for reduced latency, enhancements to LTE in unlicensed spectrum, a high-data-rate and low-energy carrier, enhancements to M2M communication, further enhancements for using multiple antennas, support for intelligent transportation systems, and enhanced support for broadcast and multicast services.

The final article elucidates a new feature in LTE-Advanced Pro that makes use of the unlicensed spectrum, known as licensed assisted access (LAA). Under LAA, licensed carriers will be able to opportunistically use unlicensed carriers to enhance the downlink performance for the user. The article discusses the built-in technologies in this new feature that will allow it to coexist with WiFi, the impact of unlicensed spectrum operation on the LTE physical layer architecture, and the scope of additional enhancements beyond LTE Release 13.

Please look forward to the next part of this Feature Topic as we characterize other evolutions envisioned for LTE.

References

[1] GSMA, "Europe Response to the European Commission Public Consultation on the Internet of Things.," http://www.gsma.com/gsmaeurope/wp-content/uploads/2012/07/GSMA-Europe-Response-EC-Consultation-IoT-10072012.pdf, July 2012.

[2] Gartner Report, "Top Strategic Predictions for 2016 and Beyond: The Future Is a Digital Thing," www.gartner.com, Oct. 2015.

Biographies

Robert W. Heath Jr. is a Cullen Trust Endowed Professor in the Department of Electrical and Computer Engineering at the University of Texas at Austin and a member of the Wireless Networking and Communications Group. He received his Ph.D. in electrical engineering from Stanford University. He is a co-author of the book *Millieter Wave Wireless*. His current research interests include millimeter-wave for 5G, cellular system analysis, communication with low-resolution ADCs, and vehicle-to-X systems.

Michael Honig (mh@eecs.northwestern.edu) is a professor in the Department of Electrical Engineering and Computer Science at Northwestern University. He received his B.S. degree in electrical engineering from Stanford University in 1977, and his Ph.D. degree in electrical engineering from the University of California, Berkeley, in 1981. Prior to joining Northwestern he worked at Bellcore in the Systems Principles Research Division. His recent research has focused on resource allocation for wireless networks and spectrum markets.

Satoshi Nagata received his B.E. and M.E. degrees from Tokyo Institute of Technology, Japan. He joined NTT DoCoMo, Inc., and worked on the research and development of wireless access technologies for LTE and LTE-Advanced. He is currently working for 5G and 3GPP standardization. He has contributed to 3GPP for many years, and contributed to 3GPP TSG-RAN WG1 as a Vice Chairman. He has been the Chairman of 3GPP TSG-RAN WG1 since 2013.

Stefan Parkvall [S'92, M'96, SM'05] is a principal researcher at Ericsson Research, active in the area of 5G research and 3GPP standardization. He received his Ph.D. degree from the Royal Institute of Technology in 1996, served as an IEEE Distinguished Lecturer 2011–2012, and co-authored several popular books such as *4G-LTE/LTE-Advanced for Mobile Broadband*. He received the Ericsson Inventor of the Year award and the Swedish government's Major Technical Award for contributions to HSPA, and was nominated for the European Inventor Award for contributions to LTE.

Anthony C. K. Soong [S'88, M'91, SM'02, F'14] (anthony.soong@huawei.com) is the chief scientist for Wireless Research and Standards at Huawei Technologies Co. Ltd., in the United States. His research group is active in the research, development, and standardization of the next generation cellular system. He has published numerous scientific papers and has over 90 patents granted or pending. He received his Ph.D. from the University of Alberta, and 2013 IEEE Signal Processing Society Best Paper Award and 3GPP2 2005 Award of Merit.

An Overview of 3GPP Enhancements on Machine to Machine Communications

Alberto Rico-Alvariño, Madhavan Vajapeyam, Hao Xu, Xiaofeng Wang, Yufei Blankenship, Johan Bergman, Tuomas Tirronen, and Emre Yavuz

The authors present an overview of several features included in 3GPP to accommodate the needs of M2M communications, including changes in the physical layer such as enhanced machine type communications, and new MAC and higher-layer procedures provided by extended discontinuous reception.

Abstract

The broad connection of devices to the Internet, known as the IoT or M2M, requires low-cost power-efficient global connectivity services. New physical layer solutions, MAC procedures, and network architectures are needed to evolve the current LTE cellular systems to meet the demands of IoT services. Several steps have been taken under the 3GPP to accomplish these objectives and are included in the upcoming 3GPP LTE standards release (3GPP Release 13). In this tutorial article, we present an overview of several features included in 3GPP to accommodate the needs of M2M communications, including changes in the physical layer such as enhanced machine type communications, and new MAC and higher-layer procedures provided by extended discontinuous reception. We also briefly discuss the narrowband IoT, which is in the development stage with a target completion date of June 2016.

Introduction

In a largely connected world, the amount of devices that access the Internet is increasing year by year. During the last decade, the increase in mobile traffic has mainly been caused by the global adoption of smartphones and the corresponding applications, which have caused the cellular networks to move from voice-centered to data-centered services. These applications require high data rate, global access to the Internet, and seamless mobility, which have been the main driver of cellular standards in the past.

In many cases, the development of *smart* devices and services, such as smart grid, smart cities, sensor networks, wearable devices, connected homes, and connected cars, pose a new set of requirements not currently supported or optimized by Long Term Evolution (LTE) cellular systems, which has the primary focus of mobile broadband (MBB) communications. The support of machine-to-machine (M2M) communications is one of the major requirements for next-generation networks [1]. Some of the key requirements of M2M communications are listed below.

Cost Reduction: In the current smartphone market, the price of the communication unit (e.g., the cellular modem) is only a small part of the overall device, which makes cost reduction much less important than other aspects, such as high peak data rate and spectral efficiency. For M2M, the cost of the communication unit has to be drastically reduced to be integrated within other types of devices, such as wearable devices (e.g., activity trackers, heart rate sensors), utility meters (water, gas, or electric), alarms, and other types of sensors. Various cost reduction techniques have been considered by the Third Generation Partnership Project (3GPP), including reduced computational complexity (e.g., reducing the bandwidth of the device or the supported transmission modes), reduced data rate, single antenna support, and half duplex operations.

Reduced Power Consumption: In many metering or sensor network applications, it is desirable to deploy battery operated devices targeting years of operation. For example, a utility company may want to collect metering information from their clients by installing a cellular modem in the metering device and having this information transmitted to a central server periodically with a duty cycle of several hours or days. Once deployed, it is expected to operate these devices over many years without the need to change batteries or redeployment. Similarly, it is critical to reduce power consumption for wearable and other tracking devices.

Enhanced Coverage: Many devices targeting M2M applications may experience poor signal reception conditions. For example, metering or alarm devices may be deployed in basements or concrete structures, which significantly increases the path loss between the transmitter and receiver. In order to reach these kinds of devices, M2M communications may require a 15–20 dB coverage enhancement with respect to regular cellular services.

There are several proprietary technologies in the so-called low-power wide-area (LPWA) family targeting Internet of Things (IoT) applications with extremely low throughput and operating in unlicensed spectrum [2]. In 3GPP, there has been an effort to enable IoT services by standardized solutions in cellular networks and to reuse the existing infrastructure as much as possible. LTE Release 12 introduced some initial features to meet the requirements driven by IoT applica-

Alberto Rico-Alvariño, Madhavan Vajapeyam, Hao Xu, and Xiaofeng Wang are with Qualcomm Inc.;
Yufei Blankenship, Johan Bergman, Tuomas Tirronen, and Emre Yavuz are with Ericsson.

0163-6804/16/$25.00 © 2016 IEEE

tions. A new user equipment (UE) category (Category 0) [3] with reduced peak data rate, half duplex operation with relaxed RF requirements, and a single receive antenna was defined to reduce the baseband and RF complexity of the UE. From the higher-layer perspective, power saving mode (PSM) [4] was adopted to allow a UE to drastically reduce power consumption for applications with delay-tolerant mobile-originated (MO) traffic in order to achieve years of battery life. In Release 13, additional improvements were introduced to drive down the cost and power consumption further. In this article we provide a high-level overview of the physical layer enhancements introduced in enhanced machine-type communications (eMTC), and the medium access control (MAC) and higher-layer improvements brought by extended discontinuous reception (eDRX). In addition, we also briefly summarize the work on the narrowband IoT (NB-IoT), which started in September 2015 with a target completion date of June 2016.

The remainder of this article is structured as follows. The next section describes the set of physical layer features introduced under eMTC. Then we briefly summarize high-level features of NB-IoT. Following that we present the higher-layer changes to support reduced power consumption under eDRX. The final section presents the conclusions.

Enhanced Machine-Type Communications

eMTC introduces a set of physical layer features aiming to reduce the cost and power consumption of UEs and extending coverage, while at the same time reusing most of the LTE physical layer procedures [5]. An eMTC UE can be deployed in any LTE evolved Node B (eNB) configured to support eMTC and can be served together with other LTE UEs by the same eNB. This allows eMTC deployment with the existing infrastructure just by applying a software update. The main features introduced by eMTC are as follows.

Narrowband Operation: The support of a wideband RF front-end and higher sampling frequencies increase the cost and power consumption of a UE. An eMTC UE follows narrowband operation for the transmission and reception of physical channels and signals, and the maximum channel bandwidth is reduced to 1.08 MHz, or 6 LTE resource blocks (RBs). This bandwidth is selected to allow the eMTC UE to follow the same cell search and random access procedures as legacy UEs, which use the channels and signals that occupy six RBs: the primary synchronization signal (PSS), the secondary synchronization signal (SSS), the physical broadcast channel (PBCH), and the physical random access channel (PRACH). The eMTC UE can be served by a cell with much larger bandwidth (e.g., 10 MHz), but the physical channels and signals transmitted or received by the eMTC UE are always contained in 1.08 MHz. A new frequency unit, called a narrowband, was defined in LTE Release 13 to accommodate this operation. A narrowband is a predefined set of six contiguous RBs in which an eMTC UE can operate. In the case of a 10 MHz channel (50 RBs), for example, 8 non-overlapping narrowbands are defined in the specification. Most of the channels of Release 12 LTE can be reused just by constraining the resource allocation to be within a narrowband. This narrowband operation is shown in Fig. 1a.

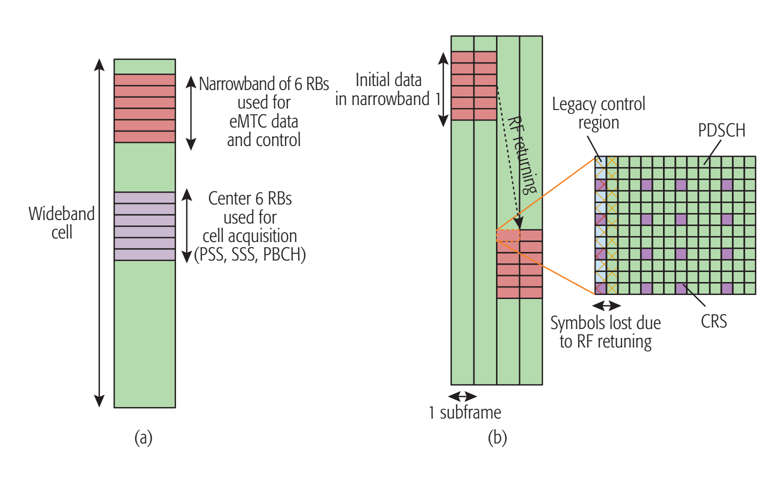

Wideband cell
Narrowband of 6 RBs used for eMTC data and control
Center 6 RBs used for cell acquisition (PSS, SSS, PBCH)
(a)
Initial data in narrowband 1
RF retuning
Legacy control region
PDSCH
Symbols lost due to RF retuning
CRS
1 subframe
(b)

Figure 1. a) Narrowband operation; b) repetition with RF retuning.

Low Cost and Simplified Operation: Many features introduced for Category 0 UEs are maintained for eMTC UEs, such as a single receive antenna, reduced soft buffer size, reduced peak data rate (1 Mb/s), and half duplex operation with relaxed switching time. New features are introduced to further reduce the cost of eMTC UEs, such as reduced transmission mode support (only transmission modes 1, 2, 6, and 9 are supported), reduced number of blind decodings for control channel, no simultaneous reception (a UE is not required to decode unicast and broadcast data simultaneously), and the aforementioned narrowband operation.

Transmission of Downlink Control Information: The legacy physical downlink control channel (PDCCH) is wideband and uses the first orthogonal frequency-division multiplexing (OFDM) symbols in a subframe, that is, control and data are multiplexed in the time domain within the same subframe. A similar structure is adopted for other control channels like the physical control format indicator channel (PCFICH) and physical hybrid automatic repeat request (HARQ) indicator channel (PHICH). A narrowband UE is not able to monitor these channels, so their functionality is replaced by new mechanisms introduced in Release 13 eMTC:

- New control channel: Instead of the legacy control channel (PDCCH), a new control channel called MPDCCH is introduced. This new control channel spans up to six RBs in the frequency domain and one subframe in the time domain. The MPDCCH is similar to enhanced PDCCH (EPDCCH), with the additional support of common search space for paging and random access.
- Handling legacy control region: In legacy LTE, the size of the control region (in number of OFDM symbols) is indicated in the PCFICH and can potentially change every subframe. In eMTC this information is semi-statically signaled in the system information block (SIB), so eMTC devices do not need to decode PCFICH.

Two modes of operation are introduced to support coverage enhancement (CE). CE Mode A is defined for small coverage enhancements, for which full mobility and channel state information (CSI) feedback are supported; CE Mode B is defined for UE in extremely poor coverage conditions, for which no CSI feedback and limited mobility are supported.

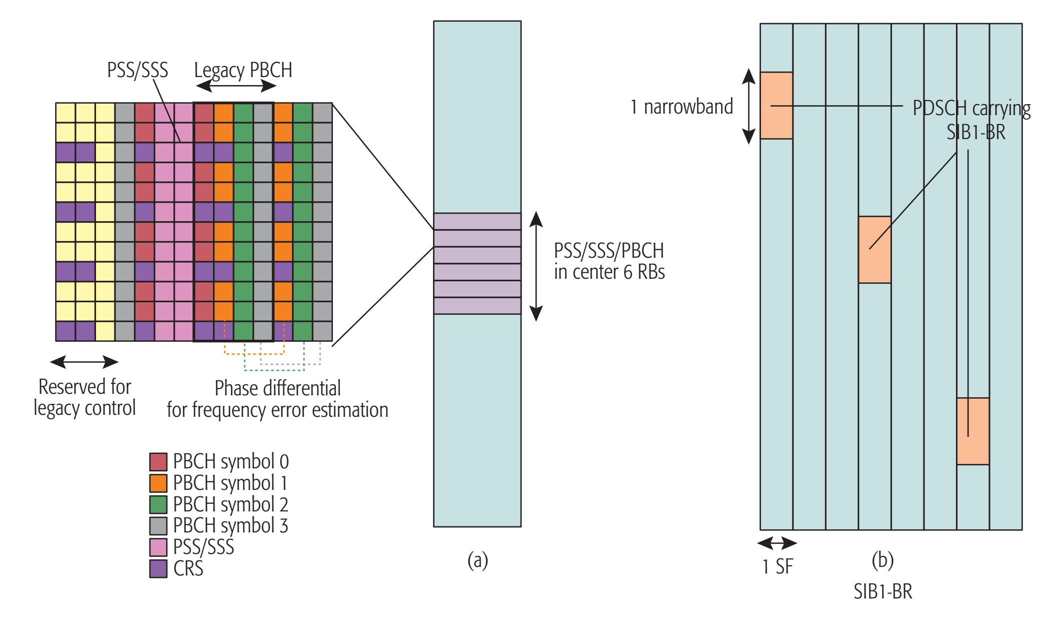


Figure 2. a) Cell search; b) system information acquisition.

• HARQ feedback for uplink transmissions: In legacy LTE this information is contained in PHICH, and retransmissions can be non-adaptive (use the same resources as the previous transmission) and are synchronous (the timing of retransmissions is fixed). In eMTC, there is no support of the PHICH, and retransmissions are adaptive, asynchronous, and based on new scheduling assignment received in an MPDCCH.

Extended Coverage: The presence of devices in extreme coverage conditions (e.g., a meter in a basement) requires the UEs to operate with much lower signal-to-noise ratio (SNR). eMTC targets 15 dB coverage enhancement with respect to Release 12 LTE, which results in 155.7 dB maximum coupling loss between transmitter and receiver. This enhanced coverage is obtained by repeating in time almost every channel beyond one subframe (1 ms) to accumulate enough energy to decode [6]. This feature is similar to uplink transmission time interval (TTI) bundling introduced in Release 8 to improve the uplink coverage for voice over IP (VoIP). The TTI bundling length, which can span 4 subframes (TTI of 4 ms) in Release 8, is extended up to 2048 subframes for the data channels in Release 13 eMTC. The following channels support repetition in eMTC: the physical downlink shared channel (PDSCH), physical uplink shared channel (PUSCH), MPDCCH, PRACH, physical uplink control channel (PUCCH), and PBCH. Two modes of operation are introduced to support coverage enhancement (CE). CE mode A is defined for small coverage enhancements, for which full mobility and channel state information (CSI) feedback are supported; CE mode B is defined for UE in extremely poor coverage conditions, for which no CSI feedback and limited mobility are supported.

Frequency Diversity by RF Retuning: Due to the narrowband RF, single antenna, and limited mobility, eMTC UEs experience limited frequency, spatial, and time diversity. In order to reduce the effect of fading and outages, frequency hopping is introduced among different narrowbands by RF retuning. This hopping is applied to the different uplink and downlink physical channels when repetition is enabled. For example, if 32 subframes are used for transmission of PDSCH, the 16 first subframes may be transmitted over the first narrowband; then the RF front-end is retuned to a different narrowband, and the remaining 16 subframes are transmitted over the second narrowband. With the assumption of a single local oscillator (LO) in the device, up to two OFDM symbols are assumed for this retuning. This narrowband operation is depicted in Fig. 1b for PDSCH repetition, where the first two OFDM symbols in the subframe after retuning are lost. Since these symbols are used for legacy control channels, the impact is limited to the loss of cell-specific reference signals (CRS) in this symbol.

In the following section we present the changes in UE operation with these new features. More precisely, we first present the new procedure for cell acquisition/initial random access and further details on data communications.

Cell Search and Initial Access

For cell search and initial access, eMTC UEs use the same signals and channels as a legacy LTE UE. The UE searches for the PSS/SSS in the center 6 RBs to obtain the cell ID, subframe timing information, duplexing mode (time-division, TDD, or frequency-division, FDD), and cyclic prefix (CP) length. There are no enhancements to PSS/SSS with the assumption that the eNB can power boost these signals to decrease the search time and power consumption of eMTC UEs in poor coverage conditions. The next step is to decode PBCH, which carries the master information block (MIB). The legacy PBCH is transmitted in the second slot of subframe 0, and for eMTC this channel is repeated in the first slot of subframe 0 and in another subframe (subframe 9 for FDD and subframe 5 for TDD). The PBCH repetition is performed by repeating the exact same constellation points in different OFDM symbols so that they can be used for initial frequency error estimation even before attempting PBCH decoding. In Fig. 2a we show the repetition pattern for subframe 0 in FDD, normal CP, and how the repeated symbols can be used for frequency error estimation. The information in the MIB is shared between eMTC UE and legacy UE, with system bandwidth, system frame number, and number of CRS antenna ports signaled to both types of UEs.

Additionally, five reserved bits in the MIB are used in eMTC to convey scheduling information about a new system information block for bandwidth reduced devices(SIB1-BR), including time and frequency location, and transport block size. SIB-BR is transmitted over PDSCH directly, without any control channel associated with it. SIB-BR remains unchanged for 512 radio frames (5120 ms) to allow a large number of subframes to be combined. In Fig. 2b we show an example of transmission of SIB-BR over the wideband LTE channel.

SIB-BR carries the basic information needed by the UE to access the system, including valid downlink and uplink subframes, maximum support of coverage enhancement, and scheduling information for other SIBs. After decoding all the necessary SIBs, the UE is able to access the cell by starting a random access procedure.

For random access, the signaling of different PRACH resources and different coverage enhancement levels is supported. This provides some control of the near-far effect for a PRACH by grouping together UEs that experience similar path loss. Up to four different PRACH resources can be signaled, each one with a reference signal received power (RSRP) threshold. The UE estimates the RSRP using the downlink CRS, and based on the measurement result selects one of the resources for random access. Each of these four resources has an associated number of repetitions for a PRACH and number of repetitions for the random access response (RAR). Thus, UE in bad coverage would need a larger number of repetitions to be successfully detected by the eNB and need to receive the RAR with the corresponding number of repetitions to meet their coverage level. The search spaces for RAR and contention resolution messages are also defined in the system information, separately for each coverage level. Note that the PRACH waveform used in eMTC is the same as in legacy LTE (i.e., based on OFDM and Zadoff-Chu sequences). Further enhancements to the PRACH waveform (e.g., [7]) may be considered in future releases if needed.

After the random access procedure is successfully completed, the UE can establish aradio resource control (RRC) connection with the eNB. The UE can be configured to be in either CE mode A or CE mode B with a UE-specific search space to receive uplink and downlink data assignments.

Data Communications

Once the RRC connection is established, the UE blindly decodes the MPDCCH in the configured search space to obtain uplink and downlink data assignments. MPDCCH is a new control channel introduced in Release 13 based on the Release 11 EPDCCH channel. An MPDCCH can be repeated in the time domain and can also use frequency hopping to improve the performance in fading channels. Unlike a legacy PDCCH, an MPDCCH uses all the available OFDM symbols in a subframe to transmit the downlink control information (DCI), so time-domain multiplexing between control and data in the same subframe is not possible. Instead, a cross-subframe scheduling rule is followed in eMTC: An MPDCCH with a last repetition in subframe N schedules a PDSCH assignment in subframe $N + 2$. DCI carried by the MPDCCH provides information on how many times the MPDCCH is repeated so that the UE knows when PDSCH transmission starts. The PDSCH assignment can be in a different narrowband, so the UE might need to retune before decoding it. For uplink data transmission, scheduling follows the same timing as legacy LTE, where an MPDCCH ending in subframe N schedules a PUSCH transmission starting in subframe $N + 4$.

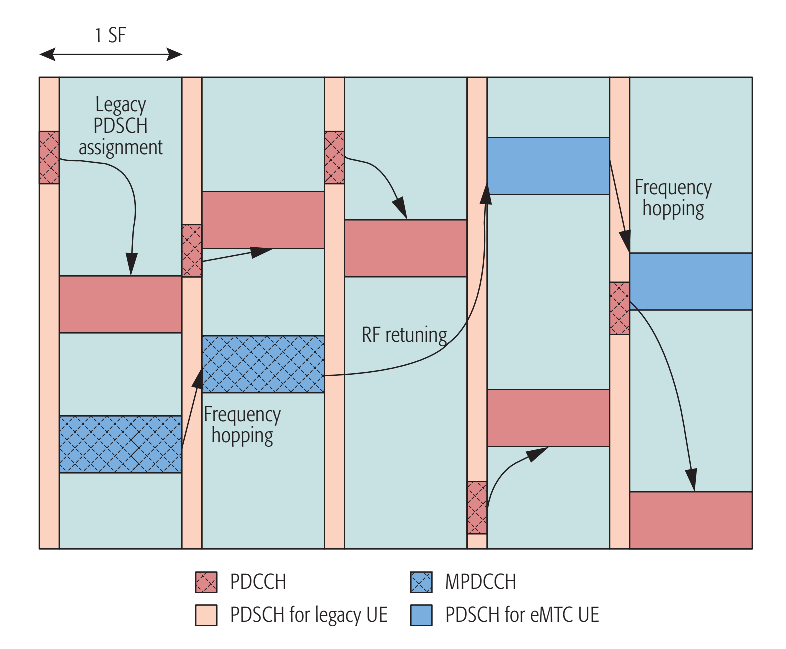


Figure 3. Scheduling timing for eMTC and legacy LTE PDSCH.

In Fig. 3 the difference in scheduling between eMTC UEs and legacy UEs is shown: legacy assignments are scheduled using the PDCCH, which uses the first OFDM symbols in each subframe. The PDSCH is scheduled in the same subframe in which the PDCCH is received. The eMTC PDSCH is cross-subframe scheduled, and a subframe is introduced between the MPDCCH and PDSCH to allow for MPDCCH decoding and RF retuning. As shown in the figure, the eNB scheduler can multiplex regular UE and eMTC UE in the same subframe just by assigning different resources and avoiding collision of the MPDCCH with the legacy PDSCH. Also, eMTC control and data channels can be repeated for a large number of subframes to be decodable in extreme coverage conditions, with a maximum number of repetitions of 256 subframes for the MPDCCH and 2048 subframes for the PDSCH.

Downlink HARQ feedback is realized by a similar mechanism as legacy LTE: a PDSCH

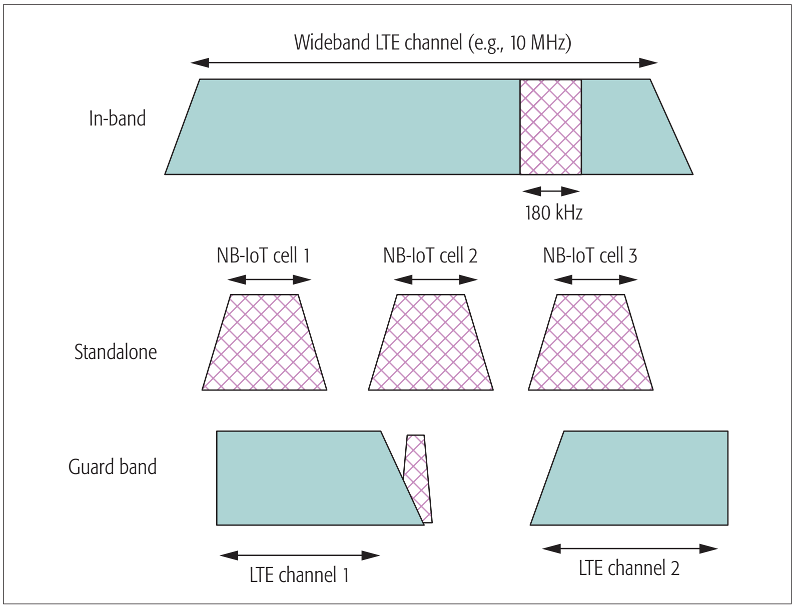


Figure 4. Different deployment modes of NB-IoT.

transmission ending in subframe *N* triggers a PUCCH transmission starting in subframe *N* + 4. Uplink HARQ is different from Release 12 LTE, as the time between retransmissions of the same HARQ process is no longer constant due to the dynamic bundling operation (bundling of a PUSCH can change in every assignment, whereas Release 8 TTI bundling is semi-statically configured by RRC signaling). HARQ retransmissions are directly triggered by receiving a new assignment over the MPDCCH. In this sense, the uplink operation in eMTC is asynchronous, following a similar procedure as downlink HARQ in previous LTE releases.

In general, eMTC introduces a wide range of features to enable cost savings and enhanced coverage while keeping great commonality with LTE, such as the reuse of most uplink and downlink physical channels. In the next section we briefly describe another technology introduced in 3GPP Release 13, which allows further bandwidth reduction to 180 kHz.

Narrowband Internet of Things

Similar to eMTC in reducing complexity and increasing coverage of cellular services, a separate work item, NB-IoT, was introduced in 3GPP for late inclusion in Release 13. Since the target completion date is June 2016, here we only provide a brief summary of what has been agreed so far for NB-IoT.

NB-IoT further decreases the bandwidth requirements compared to eMTC to 180 kHz. This narrowband bandwidth allows the device complexity to be further reduced at the expense of decreased peak data rate (around 50 kb/s for uplink and 30 kb/s for downlink). Furthermore, NB-IoT UEs only support limited mobility procedures. Thus, NB-IoT targets use cases with reduced mobility and very low data rate (e.g., metering devices) with the possibility of reusing GSM or LTE spectrum, while eMTC can cover applications with higher data rate and mobility requirements (e.g., wearables). For Release 13, both TDD and FDD are supported by eMTC, while only FDD is supported by NB-IoT (compatible with future TDD inclusion).

Unlike eMTC, which is always operated within an LTE spectrum, NB-IoT is designed to support three different deployment scenarios.

In-Band Operation: Similar to eMTC, NB-IoT can be deployed within an LTE wideband system. In this case, the narrowband comprises 1 resource block (180 kHz). For in-band operation the transmit power at the eNB is shared between wideband LTE and NB-IoT.

Standalone Operation: NB-IoT can also be deployed in a standalone 200 kHz of spectrum, for example, by *refarming* one or more GSM carriers. In standalone operation all the transmit power at the base station can be used for NB-IoT, thus increasing the coverage of these cells with respect to in-band deployment.

Guard-Band Operation: In this case, an NB-IoT cell is co-located with an LTE cell (e.g., they are served by the same eNB), but the NB-IoT channel is placed in a guard band of an LTE channel. In guard-band operation, the NB-IoT downlink can share the same power amplifier (PA) as the LTE channel, thus effectively also sharing the transmitted power.

In Fig. 4 we show a diagram of the different deployment modes of NB-IoT.

Due to the reduced bandwidth of NB-IoT, new physical channels are introduced for synchronization, broadcast information, and random access, as well as new downlink reference signals for channel estimation, tracking, and demodulation. The main NB-IoT features and differences in respect to eMTC are as follows.

Acquisition Process: eMTC and legacy LTE share the same cell search process, which includes detecting legacy PSS/SSS/PBCH. NB-IoT introduces a new set of broadcast channels and synchronization signals that use a bandwidth of 180 kHz.

Uplink Waveform: The uplink waveform of NB-IoT takes the single-carrier frequency-division multiple access (SC-FDMA) LTE uplink as a baseline, but adds some modifications on top to be more efficient in extreme coverage cases and also support lower-complexity UEs. The first change is that the minimum resource allocation is reduced from one RB to one subcarrier, thus leading to a single-tone modulation uplink transmission. In this single-tone modulation, the time-domain waveform during a symbol duration is a constant envelope sinusoid, which allows more efficient PA usage. Also, the narrower bandwidth of the uplink signals enables the multiplexing of a larger amount of UE in the same bandwidth. This increased level of multiplexing of UEs is especially useful in the case where these UEs are power limited and therefore do not benefit from being allocated a larger amount of bandwidth. Moreover, two different subcarrier spacings are allowed in NB-IoT uplink: 15 kHz (the same as legacy LTE) and 3.75 kHz (4 times lower than legacy LTE). The 3.75 kHz spacing allows for additional protection against timing errors due to the longer CP.

Downlink Transmission Schemes: A single transmission scheme based on space frequency block coding (SFBC) is supported for all physical

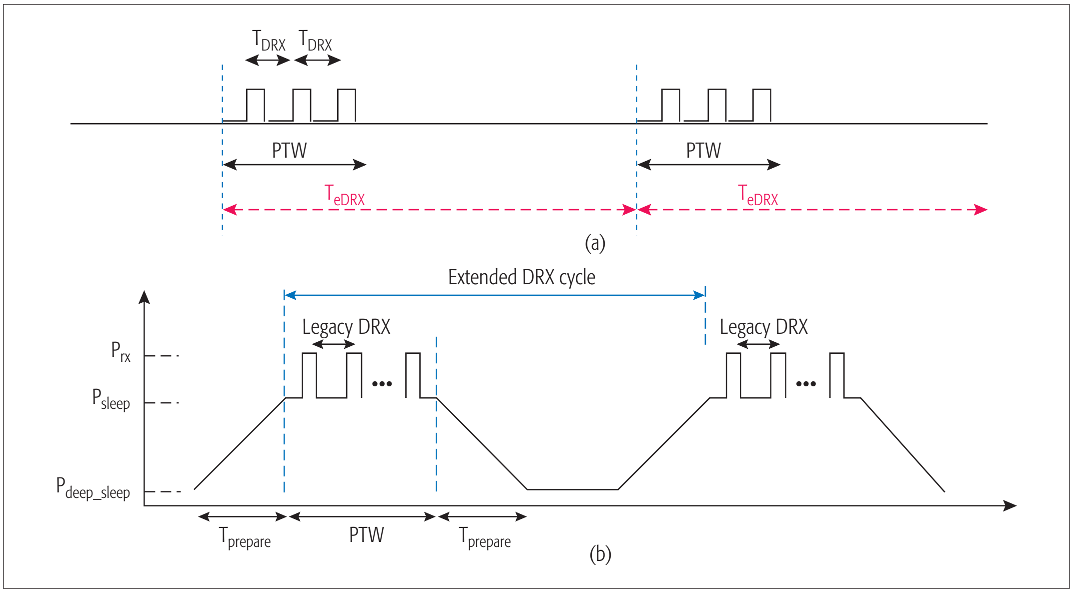


Figure 5. eDRX cycle in idle mode and resulting UE power levels.

Both eMTC and NB-IoT introduce changes in the physical layer that enable reduced device complexity and enhanced coverage, which are two of the main features required by M2M services. Although this reduced complexity also reduces the power consumption of UEs, additional modifications in the higher layer procedures are introduced to further increase the battery life of eMTC and NB-IoT UEs.

downlink channels, unlike eMTC, which supports both precoder and SFBC-based transmission schemes. One of the major changes in NB-IoT is the introduction of a new control channel based on SFBC that spans the entire subframe. A new downlink reference signal is introduced for demodulation and time/frequency tracking. Additionally, legacy LTE CRS can be used to enhance channel estimation when NB-IoT is deployed in-band.

Random Access: Unlike legacy LTE, which uses a PRACH based on Zadoff-Chu sequences, NB-IoT uses a PRACH based on single-tone transmission with hopping. A single-tone signal is used to increase the multiplexing capacity of PRACH while using a constant envelope signal, and the hopped transmission allows the eNB to estimate the round-trip delay to issue a timing advance command.

Both eMTC and NB-IoT introduce changes in the physical layer that enable reduced device complexity and enhanced coverage, which are two of the main features required by M2M services. Although this reduced complexity also reduces the power consumption of UEs, additional modifications in the higher-layer procedures are introduced to further increase the battery life of eMTC and NB-IoT UEs, which we describe next.

Extended Discontinuous Reception

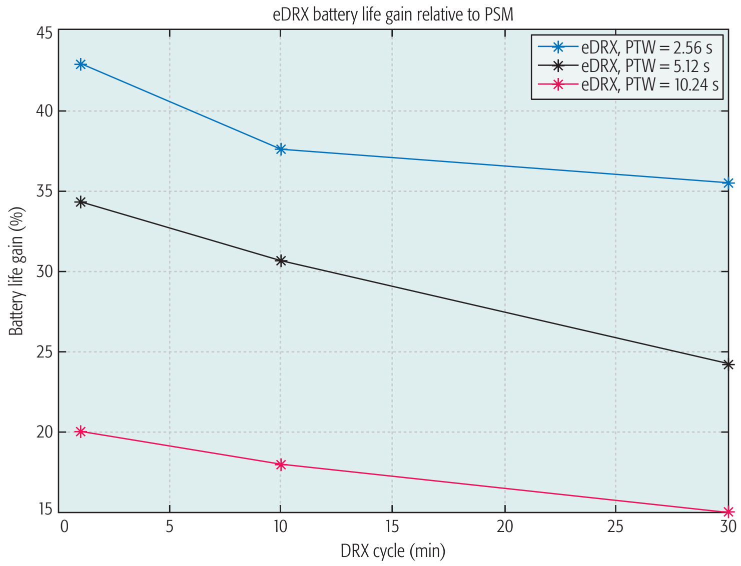
The main feature to reduce power consumption from the radio perspective is discontinuous reception (DRX). A UE configured with a DRX cycle can avoid monitoring the control channel continuously, enabling the UE to switch off parts of the circuitry to reduce power consumption. DRX is supported in Release 12 LTE with cycles up to 2.56 s. In order to efficiently support M2M communications with low duty cycle, however, the maximum DRX cycle should be increased to several minutes or even hours. A longer DRX-based mechanism for power savings enhancements has the following key advantages over PSM:

- DRX is well suited for unscheduled mobile terminated (MT) data with some requirement on delay tolerance; PSM, on the other hand, would require the UE to negotiate a periodic tracking area update (TAU) timer equal to (or slightly shorter than) the maximum allowed delay tolerance.
- A UE in DRX does not generate unnecessary signaling (resulting in power inefficiency) during periods of time when there is no data for the UE; in contrast, PSM performs TAU procedures at every TAU timer expiry, regardless of the data availability for the UE.

Extended DRX cycles are introduced in Release 13 for both idle and connected modes, thus enabling further UE power savings when the UE is not required to be reachable as frequently. For idle mode, the maximum possible DRX cycle length is extended to 43.69 min, while for connected mode the maximum DRX cycle is extended up to 10.24 s.

Since the system frame number (SFN) in LTE wraps around every 1024 radio frames (10.24 s), eDRX introduces Hyper-SFN (H-SFN) cycles to enable an extended common time reference to be used for paging coordination between the UE and the network. The H-SFN is broadcast by the cell and increments by one when the SFN wraps around (i.e., every 10.24 s). The maximum eDRX cycle corresponds to 256 hyper-frames.

A UE configured with an eDRX cycle in idle mode monitors the control channel for paging during a paging transmission window (PTW). The PTW is periodic with starting time defined by a paging hyper-frame (PH), which is based on a formula that is known by the mobility managing entity (MME), UE, and eNB as a function of the eDRX cycle and UE identity. During the PTW, the UE monitors paging according to the legacy DRX cycle (TDRX) for the duration of the PTW or until a paging message is received for the UE, whichever is earlier. The eDRX cycle and corresponding UE power consumption levels are illustrated in Fig. 5. Note that during the idle time outside of the PTW, the UE power



UE power consumption model parameters

Parameter	Value	Comments
P_{rx}	1 unit/ms	Power consumed for RX
P_{tx}	2 unit/ms	Power consumed for TX
P_{sleep}	0.01-0.02 unit/ms	Power consumed during legacy DRX sleep mode
P_{deep_sleep}	0.0002 unit/ms	Optimized sleep state
$T_{prepare}$	500 ms	Time to transition between deep sleep and sleep states
T_{eDRX}	1 min–30 min	Extended DRX cycle
T_{DRX}	2.56 s	Legacy DRX cycle

Figure 6. eDRX idle mode battery life savings gain relative to PSM.

(Pdeep_sleep) will typically be much lower than the sleep power within the PTW (Psleep). The transition to the deep-sleep state is not instantaneous and requires some preparation time for the UE to load or save the context into non-volatile memory. Hence, in order to take full advantage of power savings in deep-sleep state, the eDRX cycle (TeDRX) should be sufficiently long and the PTW as small as possible.

Figure 6 compares the battery consumption performance of devices configured with eDRX and PSM for various eDRX/TAU cycles based on the model parameters given below. The performance measure is provided as battery life gain (percent) that a device configured with eDRX can achieve over a device configured with PSM when eDRX/TAU cycles are on par. The devices are assumed to perform two MT transactions per day, and otherwise stay in deep sleep when not in PTW (if configured with eDRX) or performing TAU followed by an "active time" (if configured with PSM). It can be observed that significant power savings of up to 43 percent can be achieved when using eDRX, especially for a small PTW. When the eDRX cycle is increased, the delay tolerance also increases, and the gap between eDRX and PSM becomes smaller as expected.

Conclusions and Future Study

The support of M2M communication in cellular networks requires the introduction of new features to enable low cost, low power, and enhanced coverage. In Releases 12 and 13, several new features have been added to LTE, such as physical layer changes to reduce the UE complexity and increase coverage, and higher-layer procedures to reduce the power consumption of devices. In this article we have provided a high-level overview of the new features introduced in 3GPP Release 13: eMTC, NB-IoT, and eDRX.

Future evolution of 3GPP standardization activities related to M2M technologies may include system capacity and user throughput improvements, congestion and overload control in connected mode, position location, as well as broadcast/multicast support.

Acknowledgments

The authors would like to acknowledge Peter Gaal, Wanshi Chen, Miguel Griot, Ali Fakoorian, Onkar Dabeer, Yongbin Wei, and Durga Malladi for their valuable contributions and technical discussions.

References

[1] F. Boccardi *et al.*, "Five Disruptive Technology Directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 74–80.
[2] ETSI GS LTN 003, "Low Throughput Networks (LTN); Protocols and Interfaces v. 1.1.1 (2014-09)."
[3] 3GPP TS 36.306, "E-UTRA, UE Radio Access Capabilities (Release 12, v. 12.7.0)," 2015.
[4] 3GPP TS 23.682, "Architecture Enhancements to Facilitate Communications with Packet Data Networks and Applications (Release 12, v. 12.4.0)," 2015.
[5] 3GPP, "TR 36.888 Study on Provision of Low-Cost MTC UE Based on LTE, v. 12.0.0," 2013.
[6] R. Ratasuk, N. Mangalvedhe and A. Ghosh, "Extending LTE Coverage for Machine Type Communications," *Proc. IEEE 2nd World Forum on Internet of Things*, Milan, Italy, 2015.
[7] M. Kasparick *et al.*, "Bi-Orthogonal Waveforms for 5G Random Access with Short Message Support," *Proc. 20th Euro. Wireless Conf.*, Barcelona, Spain, 2014.

Biographies

Alberto Rico-Alvariño received his Ph.D. degree in electrical engineering from the Universidade de Vigo, Spain, in 2014. He joined Qualcomm Inc.'s (San Diego, California) Corporate R&D division in 2014 to work in standardization and development related to LTE-Advanced. He is actively involved in the standardization of eMTC and NB-IoT in LTE Release 13, among other topics.

Madhavan Vajapeyam received his B.S. degree in electrical engineering from the Universidade Federal da Paraiba, Campina Grande, Brazil, in 2000, and his M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, in 2002 and 2007, respectively. He joined the Corporate R&D division of Qualcomm in 2007, and has since been actively involved in the development and standardization of 3GPP LTE and LTE-Advanced technologies. His research interests include MAC and upper layer protocols for heterogeneous networks and machine-type communications.

Hao Xu received his B.S and M.S. from Moscow Power Engineering Institute and Technical University, Russia, in 1994 and 1996, respectively. He received his Ph.D. from Virginia Tech in 2000. During his Ph.D. research, he pioneered the millimeter-wave propagation research at 38 GHz and 60 GHz with Professor Theodore Rappaport. He received the IEEE Communications Society Steve Rice Award in 1999 with Dr. G. Durgin and Dr. T. Rappaport. From 2000 to 2003, he worked at Bell Labs' Wireless Communication Research Lab, where the first MIMO system (BLAST) was invented. His research led to one of the first outdoor MIMO channel capacity evaluations, and the joint 3GPP/3GPP2 spatial channel model (SCM). In 2003, he received the Bell-Labs President Gold Metal Award. Since 2003, he has being working at Qualcomm R&D, where his main research focus is on wireless communications system design. He has led various research and prototyping efforts, as well as 3GPP physical layer design topics. He is currently leading eMTC/NB-IoT research and standardization activities. He has numerous journal publications and patents, and also served a few years as an Associate Editor for *IEEE Transactions on Wireless Communications*.

Xiaofeng Wang received his B.S. degree in electronics from Wuhan University, China, in 1991, his M.E. degree in telecommunications engineering from Beijing University of Posts and Telecommunications, China, in 1994, and his Ph.D. degree from the University of Victoria, Canada, in 2002. Since 2000, he has worked in both industry and academia. He was with PMC-Sierra Inc. as a systems engineer from 2000 to 2002, with the Department of Electrical and

Computer Engineering, Concordia University, Montreal, Canada, as an assistant professor from 2002 to 2007, with Wavesat Inc., Montreal, Canada, as the lead of the system group from 2008 to 2010, and with PMC-Sierra Inc., as a DSP lead from 2011 and 2012. Since 2012, he has been with Qualcomm Technologies Inc. as a senior staff engineer. His research interests lie in communication theory and systems. Recently, he has been mostly involved in the development of several new communications standards such as IEEE 802.3bp 1000 BASE-T1 and 3GPP NB-IoT. He is a co-rapporteur of 3GPP NB-IoT work item.

Yufei Blankenship received her Ph.D. degree in electrical engineering from Virginia Tech in 2000. From 2000 tp 2007, she was with Motorola Labs, working on physical layer standardization with an emphasis on channel coding. She is currently a standards researcher with Ericsson. For 3GPP LTE Release 13, her focus was in the areas of MTC and NB-IoT.

Johan Bergman received an M.S. degree in engineering physics from Chalmers University of Technology, Gteborg, Sweden in 1997. He joined Ericsson in 1997 to work on baseband algorithm design for WCDMA. Since 2005, he has worked on physical layer standardization and RAN system design for WCDMA/LTE. Currently he is involved in standardization of MTC and NB-IoT in LTE Release 13.

Tuomas Tirronen is a senior researcher at Ericsson Research, which he joined in 2012 to do Internet of Things related research. He received his D.Sc. in communications engineering in 2010 from Aalto University. His current research interests in addition to IoT include 4G and 5G wireless access technologies, performance evaluation, and radio protocols and resources. He has been supporting standardization work for machine type communications in LTE Release 13.

Emre A. Yavuz received his B.Sc. and M.Sc. degrees in electrical and electronics engineering from METU, Ankara, Turkey in 1995 and 1998, respectively. He was a software engineer with Alcatel in Toronto developing safety-critical real-time microprocessor firmware for embedded command, control, and communication applications in automated train systems from 1999 to 2001. He received his Ph.D. degree in electrical and computer engineering from the University of British Columbia in Vancouver in 2007. He worked as a technical consultant from 2007 to 2009 prior to joining the School of Electrical Engineering at KTH, Stockholm, Sweden, as a postdoctoral fellow. He is now a researcher at Ericsson AB, Stockholm, working with L2/L3 standardization and RAN system design. He is currently involved in the standardization of machine type communication in LTE Release 13.

LTE Evolution for Vehicle-to-Everything Services

Hanbyul Seo, Ki-Dong Lee, Shinpei Yasukawa, Ying Peng, and Philippe Sartori

The authors provide an overview of the service flow and requirements of the V2X services LTE systems are targeting. They also discuss the scenarios suitable for operating LTE-based V2X services, and address the main challenges of high mobility and densely populated vehicle environments in designing technical solutions to fulfill the requirements of V2X services.

Abstract

Wireless communication has become a key technology for competitiveness of next generation vehicles. Recently, the 3GPP has initiated standardization activities for LTE-based V2X services composed of vehicle-to-vehicle, vehicle-to-pedestrian, and vehicle-to-infrastructure/network. The goal of these 3GPP activities is to enhance LTE systems to enable vehicles to communicate with other vehicles, pedestrians, and infrastructure in order to exchange messages for aiding in road safety, controlling traffic flow, and providing various traffic notifications. In this article, we provide an overview of the service flow and requirements of the V2X services LTE systems are targeting. This article also discusses the scenarios suitable for operating LTE-based V2X services, and addresses the main challenges of high mobility and densely populated vehicle environments in designing technical solutions to fulfill the requirements of V2X services. Leveraging the spectral-efficient air interface, the cost-effective network deployment, and the versatile nature of supporting different communication types, LTE systems along with proper enhancements can be the key enabler of V2X services.

Introduction

The concept of the "connected car" has emerged recently, in which the ability to provide a new dimension of services for drivers via wireless communications is considered as one of the most distinctive designs of next generation vehicles. Vehicles wirelessly connected to other vehicles and pedestrians within proximity can identify the possibility of collisions by exchanging information such as speed and direction at their location. Also, vehicles connected to network infrastructure can communicate with an entity in charge of traffic control so that they can be informed of unknown deterministic hazards on the road or guidance on the speed and route for traffic flow optimization. Numerous activities, including research projects and field tests, to enable connected cars are ongoing in many countries.

Widely deployed Long Term Evolution (LTE) networks and user devices can provide a means to realize this many new services for connected cars with limited cost for functional upgrade. LTE has potential to support various vehicle-to-everything (V2X) services (Fig. 1) successfully because it has an air interface of high spectral efficiency and is able to support different types of communications from one-to-one to one-to-many transmissions, and from conventional uplink and downlink cellular communications to device-to-device (D2D) direct over-the-air communications. In order to respond to this evolving market potential, the Third Generation Partnership Project (3GPP) recently started developing specifications for LTE-based V2X services with a target completion by 2016~2017.

In this article, we begin with a discussion on the significance of V2X services, and then introduce up-to-date LTE standardization activities for V2X, including the scope, use cases, and service requirements work in 3GPP. We also discuss some operating scenarios under which LTE-based V2X services are expected, and address the main challenges, such as high mobility and densely populated vehicle environments, together with technical design considerations.

V2X-Related Activities Outside 3GPP

In standardization, the intelligent transportation system (ITS), which is based on V2X communication, can be utilized for safety, non-safety, and infotainment purposes. The European Telecommunications Standards Institute (ETSI) has defined safety messages [1], which are divided into two types: cooperative awareness messages (CAMs) [2] and decentralized environmental notification messages (DENMs) [3]. CAMs are periodic messages with, for example, a frequency of 10 Hz and maximum latency of 100 ms transmitted to interchange vehicle status among vehicles in close vicinity. It is noted that in spite of the periodic nature of CAMs, the size of each message can change in time because a relatively stable information component such as the device certificate can be transmitted less frequently. DENMs are used for road hazard warnings to warn road users of dangerous events. Dedicated short-range communications (DSRC) has been developed as a standard for V2X communication, which relies on the physical and medium access layer technologies of IEEE 802.11 such as carrier sense multiple access with collision avoidance [4, 5].

In the automotive industry, automakers established the Crash Avoidance Metrics Partner-

Hanbyul Seo and Ki-Dong Lee are with LG Electronics; Shinpei Yasukawa is with NTT DoCoMo; Ying Peng is with CATT; Philippe Sartori is with Huawei.

0163-6804/16/$25.00 © 2016 IEEE

ship (CAMP) Consortium in 2001, focused on addressing the technical challenges with vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I). Several projects have been conducted in CAMP, such as Vehicle Safety Communications (VSC) [6], automated vehicle research, and the Vehicle Infrastructure Integration Consortium (VIIC). In Europe, the Car 2 Car Communication Consortium (C2C-CC) was established in 2007 [7]. C2C-CC is dedicated to increasing road traffic safety and efficiency by means of cooperative ITS (C-ITS). C2C-CC supports the creation of European standards for communicating vehicles spanning all brands.

In the government and regulatory bodies, there has been growing involvement over the past years to advance ITS. The U.S. Department of Transportation's National Highway Traffic Safety Administration (NHTSA), a federal agency authorized to oversee motor vehicle safety, together with the automotive industry and academic institutions have been researching V2V for more than a decade. In August 2014, NHTSA published an Advance Notice of Proposed Rulemaking and a supporting research report on V2V readiness that explore technical, legal, and policy issues for V2V applications. From 2012 to 2014, a safety pilot involving approximately 2800 cars, trucks, and buses from different vendors was conducted as a joint effort involving government, industry, and academia.

Standardization for V2X Services in 3GPP

The aforementioned V2X-related activities made outside 3GPP have played an essential role in motivating V2X study and normative work in 3GPP since 2014. Those activities mostly include sensor/actuator and application-layer V2X message design; however, one of the most critical problems to make the associated technology feasible in the market is how to communicate better over a wide geographic area in a more cost-effective manner; that is, the capital/operational expenditures (CAPEX/OPEX) for deployment/operation of infrastructure equipment. This is how 3GPP can provide a means for better communication (e.g., more reliable, with low latency) to make V2X more useful and cost-effective in reality. Under the ongoing standardization work, 3GPP is liaising with the other organizations to inform the outcome of the 3GPP activity as well as to solicit input [e.g., 8].

Since its first standardization, Release 8 in 2008, LTE has continued to evolve over several releases. Such evolution not only includes uplink and downlink enhancements to one-to-one communications between user equipment (UE) and base stations (also known as evolved NodeBs, or eNodeBs), but also covers other types of wireless communications. LTE supports one-to-many communications via downlink transmissions from a single cell by using single-cell point-to-multipoint (SC-PTM) transmission or from multiple cells by using multimedia broadcast multicast services (MBMS). Additionally, LTE provides a communication link called a sidelink, also known as D2D direct communications, whereby a UE directly transmits data, including data for one-to-many communications, to other UEs in its proximity without routing via an eNodeB. The one-to-many communication mechanisms in LTE are useful building blocks for V2X services, especially for safety-related services where the same message needs to be sent to multiple UEs.

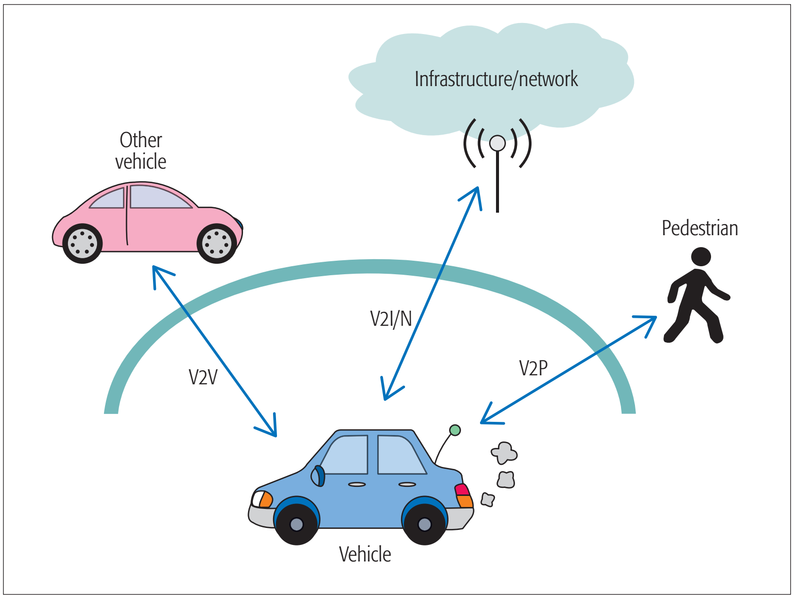


Figure 1. Types of V2X services.

In response to the increasing demand for vehicular communications, 3GPP has started standardization activities for LTE-based V2X services. As a first step, Technical Specification Group (TSG) Services and System Aspects (SA) Working Group 1 (WG1), which is responsible for defining use cases and requirements for new services and features, recently completed a study item on LTE support for V2X [9] and has started the corresponding specification work. In accordance with this outcome, TSG Radio Access Network (RAN) is conducting a study on the feasibility and identification of necessary enhancements to the LTE air interface and protocol [10]. TSG SA WG2, responsible for the overall architecture, also recently launched a study on the main functions, entities, and network interconnections required to support V2X services [11]. The 3GPP activities include V2V, vehicle-to-pedestrian (V2P), and V2I/N as target services, and cover all uplink, downlink, and sidelink communications. After the studies and corresponding specification work are finalized in 2017 as part of Release 14, a full set of technical enablers, from the air interface and protocols to the service requirement and management functionalities, will be available to support V2X services in LTE.

3GPP also recently started studies on new services that will be enabled by the new generation radio communication technology, so called fifth generation (5G), targeting specifications beyond Release 14 [12]. In this study, enhanced V2X is one of the five service categories, and the following use cases are under initial consideration:

- Autonomous driving use cases, which require very rigorous reliability (nearly 100 percent), very low end-to-end latency (e.g., a few milliseconds), and very high data rates

	Parameters					
Scenarios	Effective distance	Absolute speed of a UE	Relative speed between 2 UEs	Maximum tolerable latency	Minimum radio layer message reception reliability	(Example) cumulative transmission reliability
#1 suburban/major road	200 m	50 km/h	100 km/h	100 ms	90%	99%
#2 freeway/motorway	320 m	160 km/h	280 km/h	100 ms	80%	96%
#3 autobahn	320 m	280 km/h	280 km/h	100 ms	80%	96%
#4 NLOS/urban	150 m	50 km/h	100 km/h	100 ms	90%	99%
#5 urban intersection	50 m	50 km/h	100 km/h	100 ms	95%	–
#6 campus/ shopping area	50 m	30 km/h	30 km/h	100 ms	90%	99%
#7 imminent crash	20 m	80 km/h	160 km/h	20 ms	95%	–

Table 1. Example parameters for V2X services in 3GPP Release 14.

(e.g., tens of megabits per second), even when the density of vehicles is very high such as in multi-lane and multi-layer road scenarios
- High Mobility Mobile Broadband use cases, ensuring V2X services are available with high priority when appropriate for safety, and making mobile broadband communication seamlessly available whenever possible
- Infotainment use cases

V2X Use Cases and Service Requirements in 3GPP Release14

The study in 3GPP SA WG1 forms the basis of transport-layer-specific service and system requirements that will allow V2X-type applications (based on standards developed by other standards development organizations such as ETSI) to operate on LTE technology. The study covers three types of V2X services to be specified in 3GPP Release 14 [9]:
- V2V: covering LTE-based communication between UEs using V2V applications.
- V2P: covering LTE-based communication between UEs supporting V2P applications, where P represents vulnerable road users including pedestrians, motorcyclists, bikers, roller skaters, and so on.
- V2I/N: covering LTE-based communication between a UE and a roadside unit (RSU), both using V2I applications. An RSU is a transportation infrastructure entity (e.g., an entity transmitting speed notifications), which is implemented in an eNodeB or a stationary UE. V2N (e.g., for traffic signal control) is also included.

Both safety and non-safety use cases are possible with each type of V2X service:
- Safety-related use cases: critical-event warning (e.g., collision warning, emergency stop warning)
- Non-safety-related use cases: supplemental services that can help drivers/passengers reap the benefits of using advanced V2X services (e.g., automated parking assistance, traffic route information support)

The Technical Report [9] being produced by the feasibility study includes a wide range of categories characterizing the service requirements:
- Authentication (how to authenticate the V2X users/UEs)
- Capacity
- Charging (how mobile operators should charge for the use of V2X service)
- Communication range (measured in response time; e.g., 4 s)
- Control
- Energy consumption (communication energy efficiency due to frequent message transfer)
- Frequency of message transmissions (e.g., 10 times per second)
- Inter-operator/country (when multiple mobile operators are involved in V2X service)
- Latency (e.g., 100 ms)
- Location (sharing of location information with an improved accuracy)
- Message size (e.g., up to a maximum of 1200 bytes, excluding security overhead)
- Message transfer (timely transfer of V2X-related messages)
- Reliability
- Security (anonymity/integrity protection)
- Speed (e.g., absolute: 160 km/h; relative: 280 km/h)

Table 1 presents the key performance parameters with the suggested values for 3GPP Release 14 V2X services. The parameter *effective distance* is greater than the range required to support time to collision of 4 s at the maximum relative speed. This allows multiple V2X transmissions in order to increase the cumulative transmission reliability. *Minimum radio layer message reception reliability* denotes the probability that the recipient gets a V2X message in the effective distance and within the *maximum tolerable latency*. The parameter *cumulative transmission reliability* denotes the probability that the application at the recipient receives the required information, assuming the application layer can operate with one received V2X message during a certain time window (e.g., a 200 ms window as shown in the example; $1 - (1 - p)^2$, where p is minimum radio layer message reception reliability). Also, LTE-based V2X is working on targeting a maximum relative speed of 500 km/h for one possible scenario, although it is not listed in Table 1.

Operation Scenarios being Considered for LTE-Based V2X Services

On the operation of LTE-based V2X, two air interfaces (cellular interface based on uplink/downlink and D2D interface using sidelink) will be jointly operated and selected according to the requirement of each V2X service. The cellular communication and D2D communication, which are part of LTE-based V2X, will introduce significant operational benefit and efficient utilization of the spectrum. This subsection briefly describes the operational scenarios of LTE-based V2X together with the spectrum aspect, which is crucial in order to operate the two air interfaces and exploit conventional LTE network infrastructure.

In general, ITS consists of four types of entity; vehicles equipped with an onboard unit (OBU), vulnerable road users like pedestrians and bicycle riders, RSUs, and central ITS servers. All the entities can communicate with each other by means of cellular-based communication or D2D-based communication. D2D-based V2X will provide low latency and short-range communication even for out-of-network coverage, while cellular-based communication is for wide-area communication with high capacity. Examples of V2X deployment and transport options are shown in Fig. 2. A major difference from DSRC and ETSI ITS [4, 13] is direct network connectivity and network controllability by means of LTE infrastructure.

The RSU is a transportation infrastructure entity that could be implemented in an eNodeB or a stationary user terminal. The RSU provides several services based on the knowledge of local topology obtained from neighboring vulnerable users, sensors (e.g., cameras, induction loops), and the central ITS server. When a limited number of vehicles are equipped with OBUs, for example, at the initial stage of V2X service launch, the RSU provides local topology information obtained by roadside sensors instead of V2V communication. If an existing eNodeB can work as an RSU, rapid growth of the V2X market might be expected. Even in the mature stage, an RSU can provide wider topology information with high reliability.

A central ITS server provides centralized control for other entities as well as traffic, road, and service information. The central ITS server could be deployed outside of the LTE network by the transportation industry (e.g., a road management authority and government bodies like the Department of Transportation). Ongoing study on mobile edge computing [14] may enable deployment within the core network, that is, Evolved Packet Core (EPC), in the future in order to reduce latency.

Figure 3 shows several scenarios of spectrum usage for LTE-based V2X. It is noted that each spectrum allocated to either cellular or D2D in the figure may include multiple carriers in order to cope with high capacity requirements for future V2X services. For cellular-based V2X, existing LTE spectrum and infrastructure can be reused to offer sufficient capacity, and existing LTE networks are operated by several operators with multiple LTE carriers in a specific region. This corresponds to scenario A in Fig. 3, and as each UE uses spectrum of its own operator for both cellular and D2D links, the same spectrum can be used for both links. In this case, it is important to consider how to provide the required quality of service (QoS) for the V2X communications across UEs belonging to different operators where tight coordination and fast data transfer may not always be assumed.

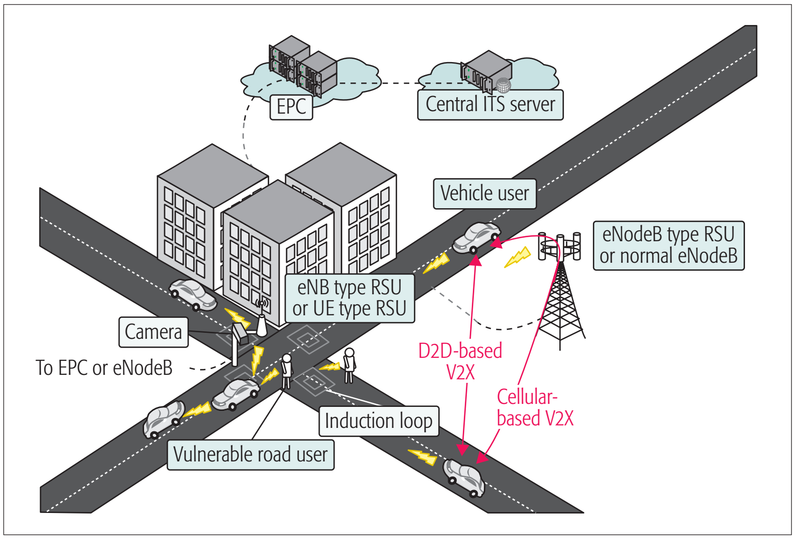
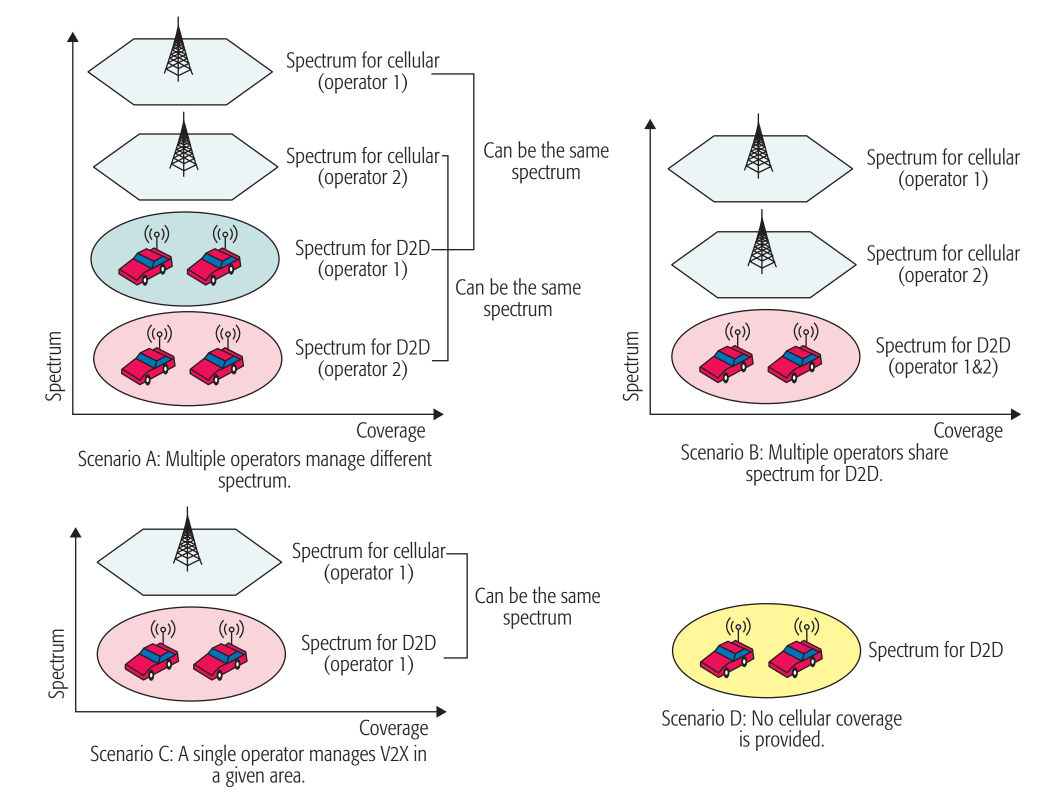


Figure 2. Examples of V2X deployment and transport options.

Depending on the frequency allocation policy, it is possible that a new dedicated spectrum is allocated to D2D-based V2X. An LTE carrier for D2D operation is not necessarily licensed to an operator. In such a case, all the D2D operation for V2X takes place in the dedicated D2D spectrum as in scenario B, and the issue of inter-operator operation is limited to the cellular link. In this case, the operator may use the cellular link for V2X services posing relatively low latency in order to account for the latency caused by the inter-operator operation, while using a D2D link for services requiring short latency and short coverage. It is noteworthy that even for D2D-based V2X, operator operation is considered for centralized control. Network control will be utilized for radio parameter optimization, radio resource allocation, congestion control, authentication and security, and so on. If no LTE coverage is provided for some areas, D2D links will be used for V2X without having such network control as in scenario D. All the parameters that would be controlled by the network can be set to predefined ones, and this may lead to relatively non-optimized operation.

If mission-critical services are supported by cellular-based V2X, dedicated spectrum for the entire V2X can have advantages in terms of capacity and QoS control. In this case, a single operator per specific area (i.e., a non-overlapping operator area) and RAN sharing operation among operators are considered as operational options with low deployment cost. As a result, the operation scenario will be in the form of scenario C in Fig. 3.

There are mainly two technical challenges in fulfilling the V2X service requirements: high vehicle speed and high UE density. It is noteworthy that UE capability may be different for vehicles and pedestrians. Higher capability and virtually unlimited battery may be assumed for UEs installed within vehicles, but the same assumption is not generally valid for pedestrian UEs.



Spectrum for cellular (operator 1)
Can be the same spectrum
Spectrum for cellular (operator 2)
Spectrum for D2D (operator 1)
Can be the same spectrum
Spectrum for D2D (operator 2)
Spectrum
Coverage
Scenario A: Multiple operators manage different spectrum.
Spectrum for cellular (operator 1)
Spectrum for cellular (operator 2)
Spectrum for D2D (operator 1&2)
Spectrum
Coverage
Scenario B: Multiple operators share spectrum for D2D.
Spectrum for cellular (operator 1)
Can be the same spectrum
Spectrum for D2D (operator 1)
Spectrum
Coverage
Scenario C: A single operator manages V2X in a given area.
Spectrum for D2D
Scenario D: No cellular coverage is provided.

Figure 3. Spectrum options for V2X operation in a given area.

Technical Challenges and Design Considerations

There are mainly two technical challenges in fulfilling the V2X service requirements: high vehicle speed and high UE density. It is noteworthy that UE capability may be different for vehicles and pedestrians. Higher capability and virtually unlimited battery may be assumed for UEs installed within vehicles, but the same assumption is not generally valid for pedestrian UEs (e.g., those installed within smartphones). Thus, consideration needs to be given to this difference in designing technical solutions.

The physical layer (PHY) design of the existing LTE system supports about 300 km/h of UE velocity at 2 GHz carrier frequency. However, PHY design for V2X faces the design objective of supporting up to 6 GHz to support a wider frequency allocation range. Also, in the D2D-based V2V scenario, the transmitter and receiver may be driven at very high velocity in opposite directions, which reaches a very high relative velocity. With such high carrier frequency and relative velocity, Doppler effects, including frequency error and inter-carrier interference, and insufficient channel estimation due to shorter coherence time, become much more serious, and current PHY design may not satisfy all scenarios.

One instance is that in the existing PHY design, two reference signals are separated by a 0.5 ms gap as shown in Fig. 4. With 500 km/h relative speed at 6 GHz spectrum, the coherence time becomes about 0.15 ms, which is smaller than the current time interval of reference signals. Consequently, the demodulation performance of the data will fall sharply because reference signals with that separation are unable to track such fast channel variations. The corresponding consideration of enhancement in the 3GPP PHY includes improving the ability to track the channel variation. Figure 4 also illustrates an example of enhanced reference signal structure where four reference signal symbols are uniformly located within a 1 ms subframe to reduce the time interval between reference signals [10]. In addition, several techniques are also under consideration by comparing the phase of the first and second half of each reference signal so that very high frequency offset can be estimated even within a single reference signal symbol.

Furthermore, high vehicle speed leads to frequent change in communication topology, which includes uplink and downlink between eNodeB and UE as well as sidelink between two UEs. Handover (i.e., change of the serving cell) is a representative example of the topology change in LTE systems, and a UE takes the new serving cell as the new reference in terms of synchronization and other communication configurations. Such change generally causes interruption in communication for some time duration. As V2X services pose tight latency and reliability requirements, V2X communications should be robust to this frequent topology change. As an example for synchronization, signals transmitted from satellites (e.g., by using GPS) can be used as the reference for sidelink, thereby allowing a reference independent of the cell change.

Compared to traditional cellular communications, V2X services are unique in terms of deployment scenarios and traffic characteristics. Vehicle density can be high, with most vehicles concentrated on a few arteries, as seen in Fig. 5.

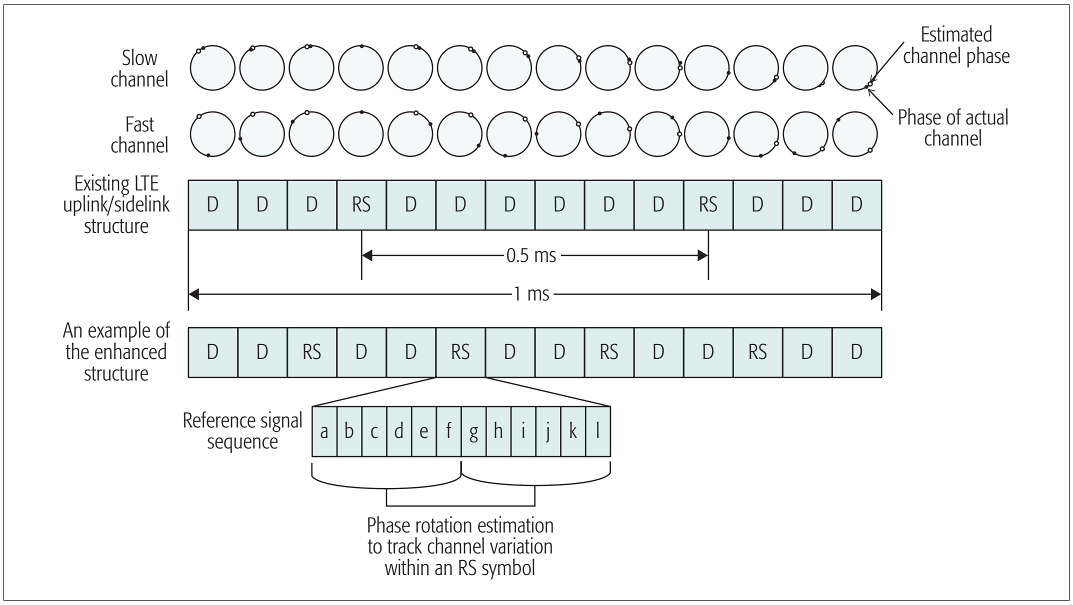

Slow channel
Fast channel
Estimated channel phase
Phase of actual channel
Existing LTE uplink/sidelink structure
D D D RS D D D D D D RS D D D
0.5 ms
1 ms
An example of the enhanced structure
D D RS D D RS D D RS D D RS D D
Reference signal sequence
a b c d e f g h i j k l
Phase rotation estimation to track channel variation within an RS symbol

Figure 4. Illustration of existing LTE uplink and sidelink channel structure and possible enhancement to track fast channel variation.

Similarly, pedestrians are concentrated in streets. In order to provide reliable safety services, it is imperative that all V2X actors transmit relatively frequently, typically every 0.1 to 1 s, using small payload sizes (less than a few hundred bytes). These traffic characteristics are very different from typical cellular communications, where a relatively small number of users are active at the same time. 3GPP is looking at several techniques in the two air interfaces to meet the demands of this challenging deployment.

The D2D interface of LTE was developed in part for public safety communications. The primary application traffic was voice, and the number of concurrent transmissions was low (tens of users in a cell area). 3GPP is investigating improvements to the D2D interface to accommodate V2X traffic. Some of the improvements being discussed include advanced resource allocation procedures to leverage the V2X traffic characteristics; for instance, most V2X traffic is periodic with a relatively predictable size. Using semi-persistent resource allocation techniques is a means to enable a large number of actors to all transmit in an efficient manner with limited signaling cost. In addition, traditional techniques such as detecting other UEs' transmission can improve the overall system performance. Given the dense environment, collision avoidance will need to be deployed. Several such techniques are currently being studied, like interference coordination, either fully autonomous between actors or with base station guidance.

Improvements for the cellular interface are also under consideration. At least for V2I/N, communication from the network needs to be considered. The transmission range typically needs to be larger than for V2V/V2P communication, and the communication is by nature point-to-multipoint. In order to accommodate this traffic demand, broadcast mechanisms are being considered with the possibility of further enhancement for the spectral efficiency and latency performance. Multi-cell broadcast based on MBMS has the benefit of reinforcing the signal strength of the message as signals from neighboring cells also act as useful ones. Single-cell broadcast using SC-PTM is beneficial in that the resource reusability of the cellular network can be exploited, and V2X messages can be efficiently multiplexed with unicast transmissions for other services.

Figure 5. An example of street deployed in the city (Seoul, Korea).

The possibility of using LTE for V2X and the expected pros and cons have been studied in several papers [e.g., 15]. Although the design for LTE-based V2X is not completed yet, its potential benefit over the existing DSRC can be summarized below.

LTE supports the frequency domain multiplexing of multiple UE transmissions in contrast to DSRC, where only one device can transmit at a time in a given channel. As a result, LTE can multiplex more UEs within limited resources without compromising each transmission's coverage, which is especially advantageous when the vehicle density is high.

Cost-Effective V2I/N: The existing LTE infrastructure, including eNBs and the core networks, can be reused with some upgrading in order to provide V2I/N services. A V2I/N service provider can save the cost of deploying new RSUs and connecting them to the network (e.g., the ITS server).

Better Coverage: LTE can provide better performance when the received signal power is weak. The receiver sensitivity is lower than that of DSRC, which means that LTE UEs can receive weak signals that are not detectable by DSRC receivers. In addition, the use of turbo code can provide better channel coding gain when compared to the convolutional code used in DSRC. Use of MBMS can be a good solution, if available, to enlarge V2X coverage.

Higher Multiplexing Capacity: LTE supports the frequency domain multiplexing of multiple UE transmissions, in contrast to DSRC, where only one device can transmit at a time in a given channel. As a result, LTE can multiplex more UEs within limited resources without compromising each transmission's coverage, which is especially advantageous when the vehicle density is high.

Robustness to Congestion: An eNodeB can allocate non-overlapping resources to different UEs in order to prevent resource collision, which is unavoidable in DSRC in a congested area. This eNodeB-based scheduling can be used for both uplink and sidelink transmissions whenever the transmitting UE is inside the network coverage. When eNodeB-based scheduling is not used, a UE can try to avoid resource collision by detecting other UEs' transmission as mentioned above, and use of semi-static allocation can be helpful in the sense that a UE can be aware of other UEs' future behavior.

Conclusions

In this article, we have discussed how LTE systems are evolving in order to support V2X services. Basic safety services such as collision warning as well as convenience services such as traffic flow optimization are identified as the first step of LTE-based V2X services. Those services can be provided in multiple operation scenarios using the D2D interface, the cellular interface, or their combination. The main challenges identified in supporting V2X services are high mobility and dense population of UEs, and LTE systems need to be enhanced so that the service requirements can be fulfilled in such a vehicular communication environment. Leveraging the spectrally efficient air interface, cost-effective network deployment, and the versatile nature of supporting different communication types, LTE systems along with proper enhancements can be a cost-effective enabler of V2X services. Furthermore, 3GPP has also started to discuss more advanced services of connected cars as the second step, and the related specification work is expected to continue for further LTE evolution and the new air interface design for 5G communications.

References

[1] ETSI TR Std 101 607, "Intelligent Transport Systems; Cooperative ITS; Release 1, v. 1.1.1, 2013; http://www.etsi.org/deliver/etsi_tr/101600_101699/101607/01.01.01_60/tr_101607v010101p.pdf.

[2] ETSI EN Std 302 637-2, "Intelligent Transport Systems; Vehicular Communications; Basic Set of Applications; Part 2: Specification of Cooperative Awareness Basic Service," v. 1.3.0, Aug. 2013; http://www.etsi.org/deliver/etsi_en/302600_302699/30263702/01.03.00_20/en_30263702v010300a.pdf.

[3] ETSI EN Std 302 637-3, "Intelligent Transport Systems; Vehicular Communications; Basic Set of Applications; Part 3: Specification of Decentralized Environmental Notification Basic Service," v. 1.2.0, Aug. 2013; http://www.etsi.org/deliver/etsi_en/302600_302699/30263703/01.02.00_20/en_30263703v010200a.pdf).

[4] ASTM E2213-03, "Standard Specification for Telecommunications and Information Exchange Between Roadside and Vehicle Systems 5 GHz Band Dedicated Short Range Communications (DSRC) Medium Access Control (MAC) and Physical Layer (PHY) Specifications," July 2003; http://www.astm.org/Standards/E2213.htm).

[5] Y. L. Morgan, "Notes on DSRC & WAVE Standards Suite: Its Architecture, Design, and Characteristics," *IEEE Commun. Surveys & Tutorials*, vol. 12, no. 4, 2010, pp. 504–18.

[6] NHTSA and CAMP, "Vehicle Safety Communications Applications (VSC-A) Final Report," 2011.

[7] CoCar Consortium, "CoCarX – ITS Services and Communication Architecture," CoCarX Deliverable D3, Oct. 2011.

[8] 3GPP, SP-150839, "LS on LTE Support for V2X Services," TSG-SA Meeting #70, Dec. 2015; http://www.3gpp.org/ftp/tsg_sa/TSG_SA/TSGS_70/Docs/SP-150839.zip).

[9] 3GPP, TR 22.885 v14.0.0, "Study on LTE Support for V2X Services"; http://www.3gpp.org/DynaReport/22885.htm.

[10] 3GPP, TR 36.885 v1.0.0, "Study on LTE-Based V2X Services"; http://www.3gpp.org/DynaReport/36885.htm).

[11] 3GPP S2-153532, "New SID on Architecture Enhancements for LTE Support of V2X Services," SA WG2 Meeting #111, Oct. 2015; http://www.3gpp.org/ftp/tsg_sa/WG2_Arch/TSGS2_111_Chengdu/Docs/S2-153532.zip).

[12] 3GPP, TR 22.891 v. 2.0.0, "Feasibility Study on New Services and Markets Technology Enablers"; http://www.3gpp.org/DynaReport/22891.htm)

[13] ETSI EN 302 665 v. 1.1.1 "Intelligent Transport Systems (ITS); Communications Architecture," Sept. 2010; http://www.etsi.org/deliver/etsi_en/302600_302699/302665/01.01.01_60/en_302665v010101p.pdf).

[14] ETSI GS MEC-IEG 005 V1.1.1, "Mobile-Edge Computing (MEC); Proof of Concept Framework," Aug. 2015; http://www.etsi.org/deliver/etsi_gs percent5CMEC-IEG percent5C001_099 percent5C005 percent5C01.01.01_60 percent5Cgs_MEC-IEG005v010101p.pdf).

[15] G. Araniti *et al.*, "LTE for Vehicular Networking: A Survey," *IEEE Commun. Mag.*, vol. 51, no. 5, 2013, pp. 148–57.

Biographies

HANBYUL SEO received his B.S., M.E., and Ph.D. degrees in electrical engineering from Seoul National University, Korea, in 2001, 2003, and 2008, respectively. He joined LG Electronics in 2008 and has been working on 3GPP LTE standardization including the areas of wireless relay, CoMP, and D2D. He is Rapporteur of 3GPP V2X standard items in TSG RAN.

KI-DONG LEE [SM '07] received his degrees in OR/industrial and systems engineering from Korea Advanced Insitute of Science and Technology (KAIST). As an author of a book, book chapters, and many journal publications, he has professional experience in probability models, scheduling, protocols, and cryptography in satellite/wireless networks. He has served as a co-Guest Editor for *IEEE Wireless Communications* and as Rapporteur of 3GPP V2X, and received awards from IEEE ComSoc, IFORS/APORS, Korea Math Society, and SK Telecom. He is an elected Vice Chairman of 3GPP SA1.

SHINPEI YASUKAWA received his M.E. degree from Doshisha University, Japan, in 2009. He joined NTT DoCoMo in 2009 and has been engaged in research and development of uplink MIMO in LTE. Since 2013, he has been engaged in 3GPP LTE standardization including the area of machine-type communications and device-to-device communication. He was a recipient of the IEICE Young Engineer Award and the Active Research Award in radio communication systems from the IEICE in 2011.

YING PENG received her M.Sc and Ph.D. degrees from the University of Bristol, United Kingdom, in 2002 and 2006, respectively, and joined CATT in 2008. She has been working on 3GPP LTE/LTE-A standardization including 4G self-evaluation, HetNet, eICIC, CoMP and D2D, and ITU-R WP5D, including 4G/5G technique evaluations since 2008. She is currently co-Rapporteur of the V2X study item and V2V work item in 3GPP RAN and co-Chairman of the 5G evaluation group in ITU-R WP5D.

PHILIPPE SARTORI is with Huawei Technologies and received his engineering degrees from ENST, France, and Ecole Polytechnique, France. He has been actively involved in the standardization of LTE since 2008. He has focused on nearly all aspects of physical design and has recently been working on device-to-device technology. He is currently co-Rapporteur of the V2V work item in 3GPP RAN. He is currently focusing on connected/autonomous vehicles and their implications on wireless communication

Advances and Challenges toward a Scalable Cloud Radio Access Network

Congmin Fan, Ying Jun (Angela) Zhang, and Xiaojun Yuan

ABSTRACT

With centralized processing, cooperative radio, real-time cloud computing, and clean infrastructure, C-RAN is a "future-proof" solution to sustain the mobile data explosion in future wireless networks. The technology holds great potential in enhancing LTE with the necessary capability to accommodate the unprecedented traffic volume that today's wireless cellular system is facing. However, the high density of RRHs in C-RANs leads to severe scalability issues in terms of computational and implementation complexities. This article discusses the challenges and recent developments in the technologies that potentially address the scalability issues of C-RANs. In particular, we focus on the collaborative signal processing, resource management, and green architecture of C-RAN systems. This article is a humble attempt to draw the attention of the research community to the following important question: how to leverage the revolutionary architecture of C-RAN to attain unprecedented system capacity at an affordable cost and complexity.

INTRODUCTION

The dramatic increase of smart mobile devices and wireless applications has led to an explosive growth in wireless data traffic. To meet the increasing traffic demand, a revolutionary wireless cellular architecture, referred to as the cloud radio access network (C-RAN), has emerged as a promising solution. A C-RAN consists of three key components:

- Distributed remote radio heads (RRHs) at remote sites of cells
- A pool of baseband units (BBUs) in a data center cloud
- A high-bandwidth low-latency optical transport network connecting the BBUs and RRHs

The key distinction of C-RANs from traditional base station (BS) systems is that the radio function units (the RRHs) are separated from the baseband processing units, and the latter are migrated to a centralized data center. This keeps RRHs lightweight, thereby allowing them to be deployed in large numbers of small cells at low cost. Meanwhile, centralized processing opens up new possibilities for significant system capacity enhancement and cost reduction through flexible interference management, dynamic spectrum reuse, collaborative radio technology, and so on. As such, the C-RAN has been recognized as a "future-proof" architecture that enables the implementation of the key features of Long Term Evolution (LTE) and LTE-Advanced, such as carrier aggregation and coordinated multipoint (CoMP).

In a C-RAN, the virtual BSs work together in a large physical BBU pool, and thus are allowed to easily share the signaling, traffic data, and channel state information (CSI) of active users in the system. On one hand, this enables tight BS coordination, including joint signal processing, scheduling, radio resource management, and load balancing, so as to greatly enhance the system capacity. On the other hand, the complexity and cost of tight coordination of all BSs may increase substantially when the network size becomes large. Indeed, the preliminary C-RAN technology can already support around 10 km separation between the BBU pool and RRHs, covering 10–1000 RRH sites. It is not hard to imagine that the size of the network will grow even larger with the advances of radio over fiber (RoF) and data center technology. With such a large network size, the computational and operational complexity of current distributed antenna systems (DASs) and multi-cell coordination schemes will become prohibitively high. Thus, it is of utmost importance to design *scalable* cooperative schemes for C-RANs, where *scalable* means:

- The computational and implementation complexity grows at the the same rate (i.e., linearly) with the network size.
- The performance is not substantially degraded compared to that of the full-scale cooperation.

In this article, we discuss the scalability issues in C-RANs from the following three aspects: signal processing, resource management, and problems related to the C-RAN architecture. Specifically, coordinated signal processing requires the knowledge and processing of large channel matrices. This leads to high channel estimation overhead to estimate the channel matrix, and high computational complexity to process the channel matrix. In this article, we introduce several recently proposed schemes for scalable signal processing. In particular, we explore the near sparsity of the channel matrix to signifi-

The authors discuss the challenges and recent developments in the technologies that potentially address the scalability issues of C-RANs. In particular, they focus on collaborative signal processing, resource management, and green architecture of C-RAN systems. They attempt to draw attention of the research community to the following important question: how to leverage the revolutionary architecture of C-RAN to attain unprecedented system capacity at an affordable cost and complexity.

Congmin Fan and Ying Jun (Angela) Zhang are with the Chinese University of Hong Kong; Xiaojun Yuan is with Shanghai Tech University.

0163-6804/16/$25.00 © 2016 IEEE

The structure of the approximated sparse matrix is not random, but closely related to the selection policy of channel entries and the architecture of C-RAN. How to use compressed sensing to estimate the sparse matrix with a special structure is still a challenging open problem.

cantly reduce the channel estimation overhead and computational complexity. For example, as shown in [1], by exploiting the near sparsity, the complexity of the optimal linear detection can be reduced from cubic to no more than quadratic with the number of RRHs. The complexity is further reduced to linear in [2]. The scalability issue of resource management in C-RANs mainly comes from the high computational complexity of solving combinatorial optimization problems. In this article, we show that game theory, graph theory, and matching theory are potential solutions for scalable resource management in C-RANs. Last but not least, we discuss the scalability issues that are closely related to the special architecture of C-RANs. Specifically, we discuss the BBU management, RRH on/off problem, and problems caused by the finite capacity of the transport network.

The purpose of this article is to draw the attention of the research community to the scalability issues in C-RANs and, in general, large-scale collaborative wireless cellular systems. It is a humble attempt to spur new research activities in this regard.

Channel Estimation and Signal Processing

Due to centralized baseband processing, RRHs in a C-RAN can be viewed as a large-scale distributed antenna system. It is widely believed that the highest system performance of a multi-antenna system is reached when all antennas are involved in cooperative transmission and reception. The fully coordinated signal processing, however, is extremely costly in a large C-RAN due to the following two reasons: (i) high channel estimation overhead to estimate the entire channel matrix, and (ii) high computational complexity to process the large-scale channel matrix. A straightforward way to decrease the high complexity on estimation and computation is to decompose the network into small clusters and limit the cooperation inside the clusters instead of over the whole network. However, due to the inter-cluster interference, clustering would inevitably degrade the benefits of full-scale cooperation, which in turn leads to a noticeable performance loss compared to full-scale cooperation. In this section, we discuss some potential solutions that decrease the estimation overhead and computational complexity without causing significant performance loss.

Channel Estimation

Full-scale RRH coordination requires knowledge of the CSI of all users to all RRHs, which results in very high channel estimation overhead. As shown in [3], the benefit of full cooperation between transmitters is fundamentally limited by the overhead of pilot-assisted channel estimation. Thus, it is of critical importance to develop efficient estimation algorithms that can estimate the C-RAN channel with minimum channel estimation overhead and high estimation accuracy. Indeed, this problem has drawn much attention for years in multiple antenna systems, especially in large-scale multiple-input multiple-output (MIMO) systems. For example, in [4], the authors derived the optimal design of training pilots for MIMO systems. However, the result cannot be extended directly to C-RAN for the following reason. In traditional MIMO systems, the transmit antennas are co-located, and so are the receiving antennas. Thus, the path loss coefficients are the same for all antennas. In C-RANs, however, both the RRHs and users are randomly located in the network area. As a result, the path loss coefficients are significantly different among different RRH-user pairs, resulting in a severe near-far problem. As such, the design of training pilots in C-RANs is much more complicated than that in large-scale MIMO systems.

Reference [5][1] considered the near-far effect and derived the optimal pilot design for a multiuser MIMO system, where the transmit antennas are co-located, but the receive antennas (or mobile users) are randomly distributed in an area. Interestingly, it proved that instead of estimating all the channel coefficients from all users, there is an optimal subset of channel coefficients that should be estimated. The optimal subset is chosen based on the path loss coefficients to balance the system capacity and channel estimation overhead. This result sheds light on the design of training pilots in C-RANs. Notice that an RRH can only receive reasonably strong signals from a small number of nearby users, and vice versa, because of the random distribution of RRHs and users over a large area. This implies that the channel matrix is a near-sparse matrix, in the sense that a majority of entries have very small magnitudes. Thus, estimating the large channel entries only, rather than the full channel matrix, can significantly simplify the channel estimation without a noticeable capacity loss. However, the unique architecture of C-RAN gives rise to a number of fundamental issues for future research, including how to select the optimal subset of channel entries for estimation, how to design the training pilots for the estimation of all the selected channel entries with minimum overhead, and, more importantly, how to make the channel estimation complexity grow at the same rate (i.e., linearly) with the size of a C-RAN.

One potential approach to estimating a near-sparse channel matrix is the well-known technique of compressed sensing. For example, [6] proposed to use compressed sensing to handle the angular sparsity in direction of arrival in a conventional MIMO channel. Recall that the channel matrix in a C-RAN can be approximated as a sparse matrix by only selecting a subset of channel entries for estimation. In this way, even though the angular sparsity does not exist in C-RANs, compressed sensing can still be applied in C-RANs. However, the structure of the approximated sparse matrix is not random, but closely related to the selection policy of channel entries and the architecture of C-RAN. How to use compressed sensing to estimate the sparse matrix with a special structure is still a challenging open problem.

Uplink Signal Detection

Besides improving the efficiency of channel estimation, the near-far effect in C-RAN channels is also a useful characteristic to improve the computational complexity scalability of uplink signal

[1] An extended version of [5] can be downloaded from http://arxiv.org/abs/1511.08977.

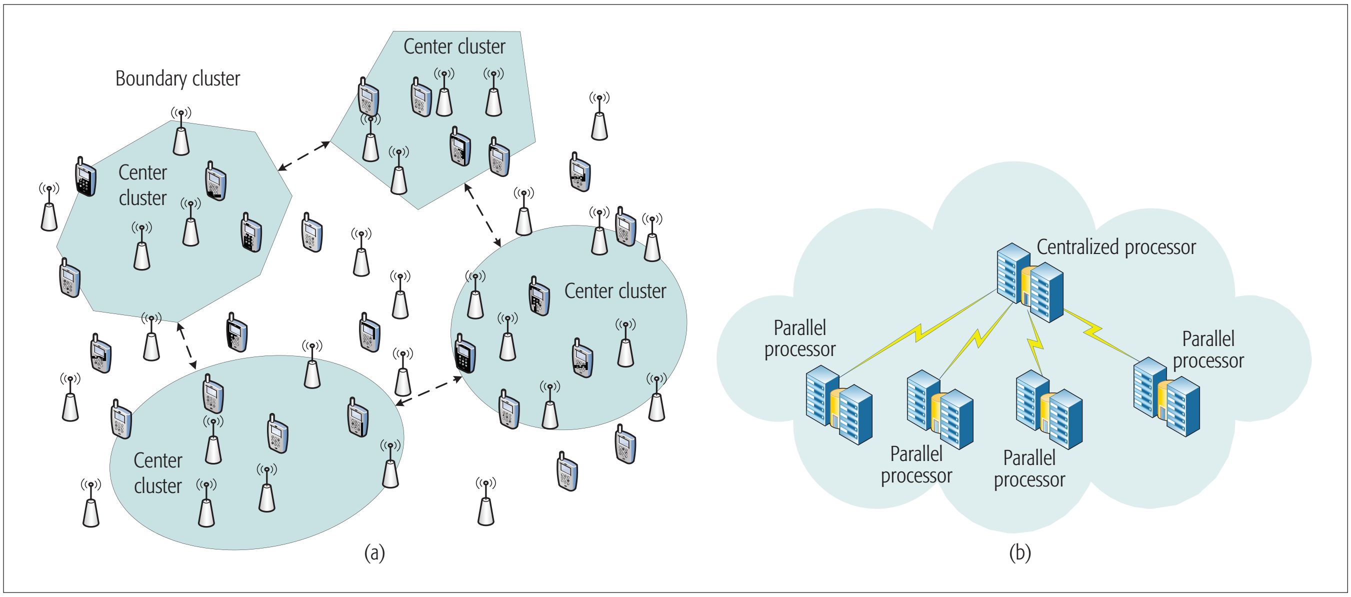


Figure 1. C-RAN architecture: a) dynamic nested clustering in a C-RAN; b) BBUs in the centralized data center.

processing. This has been explored in our recent work [1, 2]. In particular, [1] established a unified theoretical framework for dynamic clustering, consisting of the following two steps:

- Channel sparsification based on a link-distance threshold
- A detection algorithm based on dynamic nested clustering (DNC), which decomposes the centralized detection problem into problems with smaller sizes

In the first step, the channel matrix is sparsified by discarding the matrix entries if the corresponding link length (or large-scale fading in general) exceeds a certain threshold. The threshold is rigorously calculated in [1] based on the tolerance of signal-to-interference-plus-noise ratio (SINR) loss, and the location distribution of users and RRHs. Moreover, it has been proven that given a certain SINR requirement, the distance threshold does not increase with the number of RRHs/users. This means the number of non-zero channel coefficients per RRH/user does not scale with the network size. Table 1 lists the percentage of non-zero entries in the sparsified channel matrix for different SINR requirements. In the simulation, 4000 RRHs and 3000 users are uniformly located in a circular area with radius 10 km. The transmitting power allocated to each user is 80 dB beyond the noise power. We see that the non-zero entries in the channel matrix can be reduced to a very low percentage, say 0.19 percent, by compromising only 10 percent of SINR. The percentage is expected to be even lower when the network size becomes larger.

Due to the channel sparsification in the first step, the service region of each RRH is a circle centered around itself with the radius being the distance threshold. Subsequently, the second step of the algorithm divides the whole network area into multiple center clusters and a single boundary cluster (Fig. 1a). As long as the minimum distance between different center clusters is more than twice the distance threshold, there is no overlap between the service regions of RRHs from different center clusters. In this way, the center clusters can be processed independently, and they affect each other only through their interactions with the boundary cluster. As such, the signal detection complexity can be significantly reduced as it is dominated by the sizes of the clusters instead of all of the C-RAN networks. Take optimal linear detection, that is, minimum mean square error (MMSE) detection, as an example. As shown in [1], the optimal linear detection can be transformed to solving a system of linear equations defined by a doubly bordered block diagonal (DBBD) matrix (Fig. 2) based on the DNC algorithm. The diagonal blocks (except the last one) correspond to the center clusters, the cut node (i.e., the last block on the main diagonal) corresponds to the boundary cluster, and the borders capture the interaction between the center clusters and the boundary cluster. As the linear equations can be solved by processing the diagonal blocks and borders instead of manipulating the full matrix, the complexity of MMSE detection is reduced from $O(N^3)$ to $O(N^\alpha)$, where N is the number of RRHs and $\alpha \leq 2$. In this way, the computational complexity grows much more slowly with the network size.

Percentage of SINR loss (%)	1	4	7	10
Distance threshold (m)	1613	760	547	439
Percentage of non-zero entries	2.60	0.58	0.30	0.19

Table 1. Percentage of non-zero entries in sparsified channel matrix with different SINR loss requirements.

To further improve the scalability of computational complexity for uplink signal detection, [2][2] designed a randomized Gaussian message passing (RGMP) algorithm, the complexity of which grows linearly with the size of the network. In other words, the computational complexity

[2] The extended version of [2] can be downloaded from http://arxiv.org/abs/1511.09024.

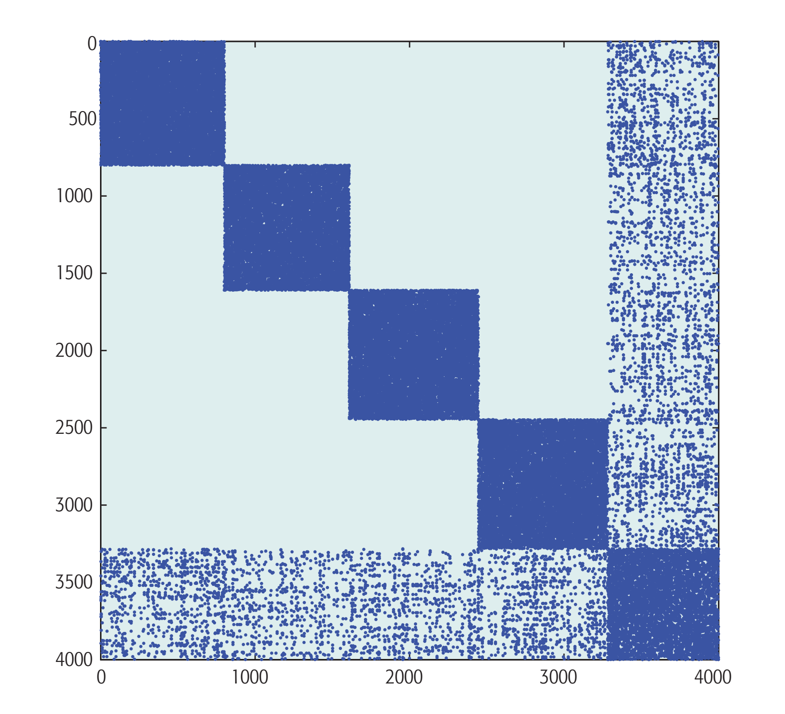


Figure 2. A doubly bordered block diagonal matrix based on the clustering given in Fig. 1a.

per RRH or user remains constant regardless of the network size, and thus perfect scalability is achieved. It can be proved that the RGMP algorithm has much better convergence than the parallel and sequential message passing algorithms. Compared to other iterative algorithms, such as the preconditioned conjugate gradient (PCG) method, the generalized approximate message passing (GAMP) algorithm, and the alternating direction method of multipliers (ADMM), the proposed RGMP algorithm has a much faster rate of convergence. As shown in Fig. 3, the number of iterations needed for convergence grows roughly linearly with the number of RRHs in both the PCG and RAMP algorithms, whereas the convergence time remains almost constant in the proposed RGMP algorithm. Moreover, Fig. 4 shows that it takes the ADMM algorithm more than 300 iterations to reduce the error to 0.01 even for a small C-RAN with only 40 RRHs, whereas the proposed RGMP algorithm converges rapidly in a few iterations.

The successful algorithms proposed in [1, 2] show the possibility of designing scalable signal processing for C-RAN. Future extensions can be envisioned, such as extension to nonlinear signal detection, joint channel estimation and signal detection, and so on.

Downlink Beamforming

Unlike the uplink case, the beamforming design in downlink involves transmit power constraints of individual RRHs. This makes the downlink beamforming more complicated. To deal with the high complexity, researchers have proposed several efficient algorithms, among which ADMM is particularly outstanding due to its distributed and parallelizable implementation. In [7], a two-stage framework based on ADMM is presented to solve large-scale convex optimization problems, such as downlink beamforming problems. In the first stage, the original problem is transformed into a standard cone programming form via matrix stuffing. In the second stage, ADMM is adopted to solve the problem in a standard form. Since ADMM can solve the problem in parallel, the computational time is significantly reduced.

Reference [8] showed that the dual of a multi-antenna downlink channel with per-antenna power constraints is an uplink channel with noises that has an uncertain diagonal covariance matrix. Then the original downlink problem can be solved by iteratively updating the noise covariance matrix and the dual uplink detection matrix. Thanks to the similarity between channel models of C-RANs and traditional multi-antenna systems, the downlink-uplink duality given in [8] also holds for C-RANs. This indicates that the downlink beamforming problem in C-RANs may be solved by existing efficient uplink algorithms, such as those proposed in [1, 2].

Considering the satisfactory performance of message passing in uplink, it is reasonable to treat message passing as one promising solution to downlink beamforming. Sohn *et al.* presented a message passing algorithm for downlink beamforming to maximize the sum throughput in a cooperative MIMO network [9]. The algorithm has a polynomial-time computational complexity and is amenable to parallel implementation. Although this work only considered small cooperative networks, it sheds light on the possibility of designing beamforming algorithms based on message passing.

Another promising scheme to reduce the complexity of beamforming design is clustering, which limits the cooperation inside each cluster. As mentioned before, since cluster-edge users suffer from severe inter-cluster interference, clustering leads to inevitable performance loss. A recent work by Ratnam *et al.* [10] proposed an interlaced clustering algorithm as a solution to the edge user problem. In interlaced clustering, multiple cluster patterns are spatially shifted replicas of each other, and operate simultaneously. Thus, an edge user on one cluster pattern is in the interior of a cluster on another cluster pattern with high probability, which ensures good throughput for each user. A number of future extensions can be envisioned, such as designing cluster patterns to improve the system performance, deriving the optimal cluster size and the optimal number of cluster patterns, reducing the complexity of beamforming inside each cluster, and so on.

Resource Management

Besides coordinated signal processing, coordinated resource management, including dynamic frequency reuse, coordinated scheduling, and RRH association, now becomes possible because of the centralized C-RAN architecture. However, the optimal frequency, time, and RRH allocation problem is a combinatorial optimization problem. The computational complexity of solving such a problem is prohibitively high, especially in a large-scale C-RAN. In this section, we present some promising schemes that may be used to avoid the combinatorial computational complexity in resource allocation problems.

Game theory has attracted much attention in wireless networks, as resource allocation problems can easily be transformed to games. In a game, RRHs/users are treated as players that aim to maximize their own utility or the overall network efficiency by choosing when, on which channel, and to whom to transmit. For example, a recent work by Bethanabhotla *et al.* [11] proposed decentralized association schemes to solve the user-cell association problem in massive MIMO based on game theory. In the article, an association game is formulated by defining the users as players and the users' throughputs as payoff functions. In the game, each user makes its own association decisions to BSs based on its own user-centric utility function. As shown in [11], such a decentralized association scheme achieved good performance with low complexity.

Another promising technique for resource allocation in C-RANs is graph theory. In [12], a graph-theoretical-based approach is proposed to solve the coordinated scheduling problem in C-RANs. A scheduling graph is constructed by setting a vertex as an association between user, BSs, and time/frequency resource block. Some vertices in the scheduling graph are connected to represent practical constraints in the original problem. For example, two users cannot be served by the same time/frequency block, or a user cannot be connected to multiple BSs. In this way, the original scheduling problem is reformulated as a maximum weight clique problem, where the weight of each vertex is the benefit of the association represented by that vertex. As shown in [12], such a maximum weight clique problem can be solved easily by efficient algorithms in graph theory.

Recently, motivated by the tractable solutions and efficient algorithmic implementations of matching theory for combinatorial problems, Gu *et al*. gave a tutorial on the use of matching theory for resource allocation in wireless networks [13]. To illustrate the concepts of matching theory, the tutorial proposed a wireless-oriented classification of matching theory. Based on the players' quotas, there are three classical types of matching: one-to-one matching, many-to-one matching, and many-to-many matching. Moreover, to capture the wireless resource allocation features, three novel classes are proposed: canonical matching, matching with externalities, and matching with dynamics. Specifically, in canonical matching, the preference of each resource/user is independent of other resources'/users' choices. In matching with externalities, the preference of each resource/user varies with other resources'/users' choices due to "peer effects," such as interference. Matching with dynamics is more complicated, in which the matching processes are sensitive to dynamics of the environment, inc;luding fast fading, mobility, and so on. With the detailed classification of matching, it is easy to transfer a resource allocation problem to a matching problem. For example, we consider a C-RAN frequency reuse problem, in which each user is assigned to one channel, and different users can transmit over the same channel. Obviously, this problem is a many-to-one matching. Since the users occupying the same channel will cause interference to each other, the problem belongs to matching with externalities. With existing techniques in matching theory, it is very likely to solve the above-mentioned matching problem with efficient algorithms.

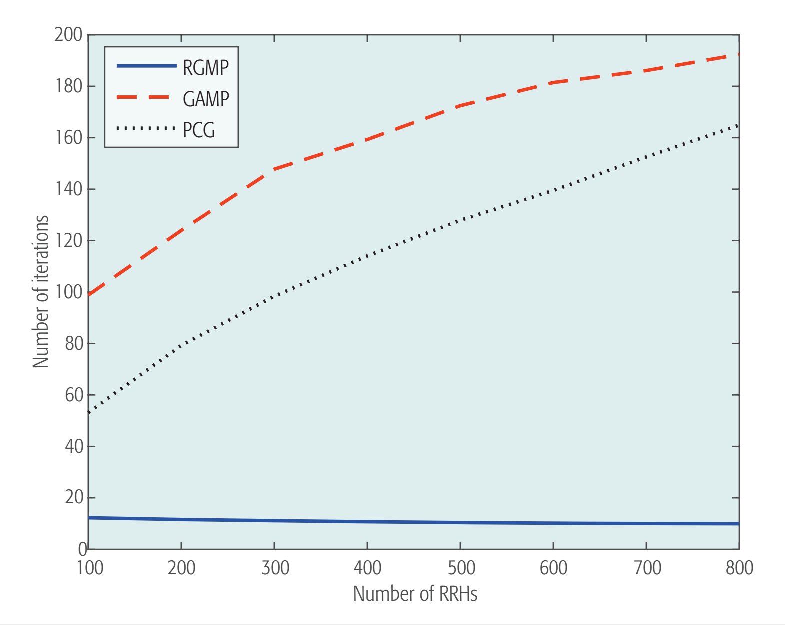


Figure 3. Convergence rate against the number of RRHs.

So far, we have only discussed some prospective techniques for efficient resource allocation but have not considered the near sparsity of the C-RAN channel matrices. The near sparsity of channel matrices can significantly simplify the above-mentioned techniques. For example, we consider the user-cell association problem presented in [11] for C-RANs instead of massive MIMO. The association lists of users can be shortened since in C-RANs, including an RRH in the association list of a far-off user is meaningless. The system capacity would not be noticeably increased by allowing RRHs to serve a far-off user since the corresponding channel coefficients are very small.

Green Architecture

The special features of C-RANs lead to some unique architecture problems, such as those below.

BBU Management: The C-RAN architecture allows the data center to dynamically adjust the workload among BBUs. Efficient algorithms for computational power allocation and workload scheduling become a necessity. Moreover, notice that the wireless traffic load is highly dynamic in time and space. Thus, when the amount of workload is small, it is energy-efficient to consolidate all workload to a subset of BBUs and turn off the idle BBUs. However, once in off state, a BBU cannot be restarted instantaneously. This would incur a service delay if the active BBUs cannot satisfy all the incoming workload. Dynamic BBU management mechanisms are needed to balance the trade-off between energy efficiency and service delay.

Limited Capacity of Transport Network: To guarantee seamless cooperation among

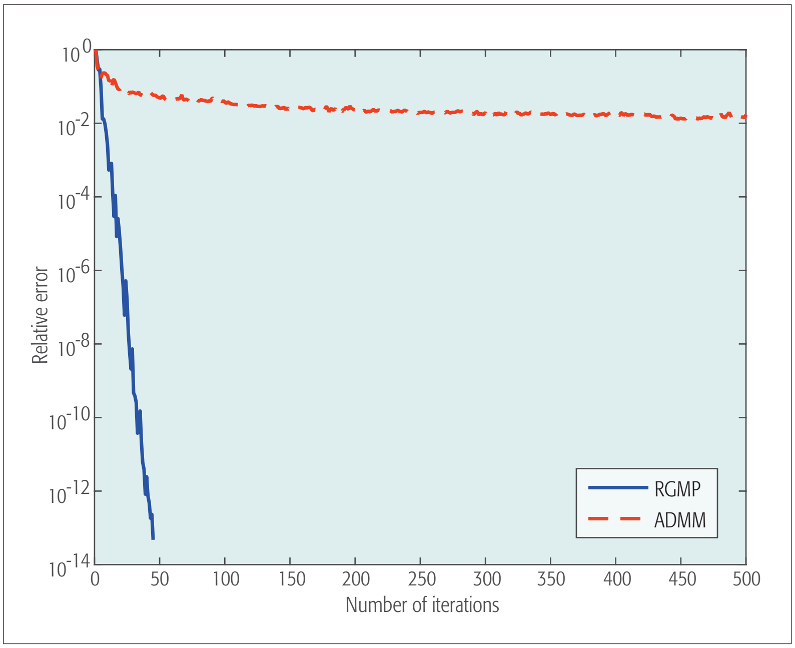


Figure 4. Relative error vs. number of iterations when there are 40 RRHs and 32 users in a C-RAN.

RRHs, data of RRHs transmit to the centralized data center from time to time. Even though high-bandwidth low-latency optical transport links are employed in C-RANs, the capacity of the transport network is still finite. This limits the amount of information that can be exchanged between the BBUs and RRHs, which implies that the cooperation among RRHs is limited. Moreover, the virtually centralized BBUs in the data center are geographically separated in general. Thus, the BBUs also need to share data over the transport network. How to fully utilize the limited backhaul capacity to maximize the cooperation and thereby optimize the system capacity is a critical problem in C-RANs.

RRH On/Off Problem: Thanks to the high density of RRHs and fluctuation of wireless traffic load, not all the RRHs need to be active all the time. For example, on weekends, it is essential to turn off some RRHs in office and industrial zones to decrease the power consumption in C-RANs. The corresponding power consumption of the transport network connected to these RRHs can also be reduced. The RRH on/off scheme is undoubtedly one of the key techniques to improve the energy efficiency in C-RANs.

It is clear that all the above-mentioned problems are highly related to the signal processing and resource management schemes. Thus, as shown in our later discussions, these problems are often jointly solved with detection, beamforming, scheduling, and so on.

In [1], a computational power allocation scheme associated with the DNC algorithm was proposed for the parallel implementation of signal processing. As shown in Fig. 1b, the operations corresponding to the boundary cluster are performed by the centralized processor, and those corresponding to the center clusters are performed by the parallel processors. The interactions between the boundary cluster and the center clusters are captured by exchanging messages between the centralized processor and the parallel processors (denoted by lighting symbols in Fig. 1b). By adjusting the cluster sizes in C-RANs, the amount of workload allocated to each processor can easily be changed. This computational power allocation scheme is amenable to different architectures of the BBU pool. Moreover, in this scheme, the number of messages to be shared is very small. For example, [1] showed that for MMSE detection, messages only need to be exchanged twice among different processors. This will decrease the traffic load of the transport network among processors.

Zhou and Yu studied the uplink of a C-RAN with limited backhaul capacity in [14]. Instead of directly sending the received signals to the BBU pool as in [1], [14] quantized the received signals before sending them to the BBU pool. This compress-and-forward scheme can significantly reduce the amount of data to be transmitted over the transport network, but also inevitably introduces quantization noises. Thus, an optimization approach was proposed in [14] to optimize the sum-rate by minimizing the quantization noise level under a backhaul capacity constraint. This work presented an efficient backhaul capacity allocation algorithm in C-RANs.

A work by Shi *et al.* formulated a power minimization problem by jointly considering RRH selection and beamforming vector design [15]. Notice that when an RRH is turned off, all the corresponding entries in the beamforming vector are set to zero. Thus, the beamforming vector is in a group sparsity structure. In this way, the problem is simplified to a group sparse beamforming problem. Efficient algorithms were designed by applying the l1/lp-norm to induce group sparsity. Reference [15] is a successful attempt to solve the RRH selection problem by group sparsity, and has motivated several follow-up works, which considered joint beamforming and RRH selection problems with backhaul capacity constraints, imperfect CSI, or some other constraints.

The above three works presented some initial ideas to improve the architectural efficiency in C-RANs. Future work on architecture problems is expected to construct a green and environment-friendly C-RAN.

Conclusions

Featuring centralized baseband processing, cooperative radio, and real-time cloud infrastructure, C-RAN has great potential to be a predominant wireless cellular architecture in next-generation wireless systems. By introducing the challenges and some potential solutions to enhance the scalability of C-RAN algorithms, this article aims to attract research efforts on designing efficient schemes for scalable signal processing, resource management, BBU management, and so on. These schemes, if successfully developed, will greatly advance the key technologies of C-RANs, and consequently contribute to the paradigm shift to LTE and LTE-Advanced in wireless networks.

Acknowledgments

This work was supported in part by the National Basic Research Program (973 Program number 2013CB336701), the National Natural Science Foundation of China (Project numbers 61471241, 61201262), and two grants from the Research Grants Council of Hong Kong under General Research Funding (Project numbers 2150828 and 2150876).

References

[1] C. Fan, Y. J. Zhang, and X. Yuan, "Dynamic Nested Clustering for Parallel PHY-Layer Processing in C-RANs," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, Mar. 2016, pp. 1881–94.

[2] C. Fan, X. Yuan, and Y. J. Zhang, "Randomized Gaussian Message Passing for Scalable Uplink Signal Processing in C-RANs," *Proc. IEEE ICC*, Kuala Lumpur, Malaysia, 2016.

[3] A. Lozano, R. Heath, and J. Andrews, "Fundamental Limits of Cooperation," *IEEE Trans. Info. Theory*, vol. 59, no. 9, Sept. 2013, pp. 5213–26.

[4] B. Hassibi and B. M. Hochwald, "How Much Training Is Needed in Multiple-Antenna Wireless Links?" *IEEE. Trans. Info. Theory*, vol. 49, Apr. 2003, pp. 951–63.

[5] C. Fan, X. Yuan, and Y. J. Zhang, "Throughput Bounds for Training-Based Multiuser MIMO Systems," *Proc. IEEE ICCCN*, Waikoloa, HI, 2016.

[6] W. U. Bajwa *et al*., "Compressed Channel Sensing: A New Approach to Estimating Sparse Multipath Channels," *Proc. IEEE*, vol. 98, no. 6, June 2010, pp. 1058–76.

[7] Y. Shi et al., "Large-Scale Convex Optimization for Dense Wireless Cooperative Networks," *IEEE Trans. Signal Processing*, vol. 63, no. 18, Sept. 2015, pp. 4729–43.

[8] W. Yu and T. Lan, "Transmitter Optimization for the Multi-Antenna Downlink with Per-Antenna Power Constraints," *IEEE Trans. Signal Processing*, vol. 55, no. 6, June 2007, pp. 2646–60.

[9] I. Sohn, S. H. Lee, and J. G. Andrews, "Belief Propagation for Distributed Downlink Beamforming in Cooperative MIMO Cellular Networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 12, Dec. 2011, pp. 4140–49.

[10] V. V. Ratnam, G. Caire, and A. F. Molisch, "Capacity Analysis of Interlaced Clustering in a Distributed Antenna System," *Proc. IEEE ICC*, London, U.K., 2015.

[11] D. Bethanabhotla *et al*., "Optimal User-Cell Association for Massive MIMO Wireless Networks," *IEEE Trans. Wireless Commun*., vol. 15, no. 3, 2015, pp. 1835–50.

[12] A. Douik *et al*., "Coordinated Scheduling for the Downlink of Cloud Radio-Access Networks," *Proc. IEEE ICC*, London, U.K., 2015.

[13] Y. Gu *et al.*, "Matching Theory for Future Wireless Networks: Fundamentals and Applications," *IEEE Commun. Mag*., vol. 53, no. 5, May 2015, pp. 52–59.

[14] Y. Zhou and W. Yu, "Optimized Backhaul Compression for Uplink Cloud Radio Access Network," *IEEE JSAC*, vol. 32, no. 6, June 2014, pp. 1295–1307.

[15] Y. Shi, J. Zhang, and K. B. Letaief, "Group Sparse Beamforming for Green Cloud-RAN," *IEEE Trans. Wireless Commun*., vol. 13, no. 5, May 2014, pp. 2809–23.

Biographies

Congmin Fan [S'14] (fc012@ie.cuhk.edu.hk) received her B.Eng. degree with first class honors in information engineering from the Chinese University of Hong Kong in 2012. She is currently pursuing her Ph.D degree at the same university. Her research interests are 5G wireless communication networks, cloud radio access networks, and optimization theory, with current emphasis on signal processing, resource allocation, and channel estimation.

Ying Jun (Angela) Zhang [S'00, M'05, SM'11] (yjzhang@ie.cuhk.edu.hk) received her B.Eng. degree in electronic engineering from Fudan University, Shanghai, China, in 2000, and her Ph.D. degree in electrical and electronic engineering from Hong Kong University of Science and Technology in 2004. Since 2005, she has been with the Department of Information Engineering, Chinese University of Hong Kong, where she is currently an associate professor. During the summers of 2007 and 2009, she was with the Wireless Communications and Network Science Laboratory at Massachusetts Institute of Technology (MIT). Her current research interests are mainly focused on wireless communications systems and smart power systems, in particular optimization techniques for such systems. She is an Executive Editor of *IEEE Transactions on Wireless Communications*. She is also an Associate Editor of *IEEE Transactions on Communications*. Previously, she served many years as an Associate Editor of *IEEE Transactions on Wireless Communications* and *Security and Communications Networks* (Wiley), and a Guest Editor of a Feature Topic in *IEEE Communications Magazine*. She has served as a Workshop Chair of IEEE ICCC 2014 and 2013, TPC Vice Chair of Wireless Networks and Security Track of IEEE VTC 2014, TPC Vice-Chair of Wireless Communications Track of IEEE CCNC 2013, TPC Co-Chair of Wireless Communications Symposium of IEEE GLOBECOM 2012, Publication Chair of IEEE TTM 2011, TPC Co-Chair of the Communication Theory Symposium of IEEE ICC 2009, Track Chair of ICCCN 2007, and Publicity Chair of IEEE MASS 2007. She was a Co-Chair of the IEEE ComSoc Multimedia Communications Technical Committee and the IEEE Communication Society GOLD Coordinator. She is a co-recipient of the 2014 IEEE ComSoc APB Outstanding Paper Award, 2013 IEEE SmartgridComm Best Paper Award, and 2011 IEEE Marconi Prize Paper Award on Wireless Communications. She received the Young Researcher Award from the Chinese University of Hong Kong in 2011. As the only winner from Engineering Science, she won the Hong Kong Young Scientist Award 2006, conferred by the Hong Kong Institution of Science.

Xiaojun Yuan [S'04, M'09] (yuanxj@shanghaitech.edu.cn) received his B.S. degree in electronic and information systems from Shanghai Jiaotong University, his M.S. degree in circuit and systems from Fudan University, and his Ph.D. degree in electrical engineering from the City University of Hong Kong in 2008. From 2009 to 2011, he was a research fellow at the Department of Electronic Engineering, City University of Hong Kong. He was a visiting scholar at the Department of Electrical Engineering, University of Hawaii at Manoa in spring and summer 2009, as well as in the same period of 2010. From 2011 to 2014, he was a research assistant professor at the Institute of Network Coding, Chinese University of Hong Kong. He is now an assistant professor at the School of Information Science and Technology, ShanghaiTech University. His research interests cover a broad range of wireless communications, signal processing, and information theory including MIMO techniques, physical-layer network coding, cooperative communications, compressed sensing, and so on. He has published over 80 peer-reviewed research papers in leading international journals and conferences, and has served on a number of technical program committees for international conferences. He was a co-recipient of the Best Paper Award of the IEEE International Conference on Communications, Sydney, Australia, in 2014.

It is essential to turn off some of RRHs in the office and industrial zones to decrease the power consumption in C-RANs. The corresponding power consumption of the transport network that are connected to these RRHs can also be reduced. The RRH on/off scheme is undoubtedly one of the key techniques to improve the energy efficiency in C-RANs.

Wireless Communication for Factory Automation: An Opportunity for LTE and 5G Systems

Bernd Holfeld, Dennis Wieruch, Thomas Wirth, Lars Thiele, Shehzad Ali Ashraf, Jörg Huschke, Ismet Aktas, and Junaid Ansari

Wireless factory automation is an application area with highly demanding communication requirements. The authors classify these requirements and identify the opportunities for the current LTE air interface for factory automation applications. Moreover, they give an outlook on the relevant design considerations to be addressed by 5G communication systems.

Abstract

The evolution of wireless communication from 4G toward 5G is driven by application demands and business models envisioned for 2020 and beyond. This requires network support for novel use cases in addition to classical mobile broadband services. Wireless factory automation is an application area with highly demanding communication requirements. We classify these requirements and identify the opportunities for the current LTE air interface for factory automation applications. Moreover, we give an outlook on the relevant design considerations to be addressed by 5G communication systems.

Introduction

In the Next Generation Mobile Network (NGMN) consortium and the Third Generation Partnership Project (3GPP), the use case of machine type communication (MTC) is divided into two main groups as massive MTC (M-MTC) and mission-critical MTC (C-MTC) [1]. While M-MTC involves a large number of low-cost devices such as sensors or meters with high requirements on coverage and energy efficiency, C-MTC targets scenarios with very low latency and high reliability requirements such as process automation, intelligent transportation systems, and smart grid, as well as factory automation.

In this article, we focus on the highly challenging factory automation applications with strict demands on latency and reliability. In this context, reliability refers to guaranteed message delivery within the required latency bound. It is typically quantified as the residual block error rate (BLER) at the physical layer (PHY) or the packet error rate (PER) at higher layers of the protocol stack. Latency is considered end-to-end (e2e) in factory automation, where one end is formed by sensors measuring data and providing it to the process logic controller (PLC). The PLC comprises the essential logic to process the collected sensor data and instructs the actuators forming the other communication end.

In recent years, using wireless technologies for factory applications has received significant attention from the automation and communication industry. This is mainly attributed to the following. First, installation and maintenance cost for cables are high as they often experience wear and tear, need additional protection and housing, and limit mobility due to interleaving. Therefore, from time to time cables have to be replaced manually, which requires intervention of trained personnel and interruption of production processes. In contrast, wireless technologies have very low installation and maintenance costs. Second, wireless technologies offer a high degree of deployment flexibility, which enables rapid realization of different production deployments, even with mobility. Finally, the shared nature of the wireless medium allows communication flexibility, where any device can communicate with any other device in its communication range.

In the scope of the KoI project [2] funded by the German Federal Ministry of Education and Research (BMBF), seven partners from industry and academia are investigating the wireless factory automation scenario. The novel communication architecture proposed in the KoI project is based on two-tier radio resource coordination as illustrated in Fig. 1. The two-tier architecture is chosen to realize a logical separation of mission-critical functionalities from generic functionalities. Note that in practical implementations, these functionalities could potentially be integrated within a single entity. On the first tier, the global radio coordinator uses Long Term Evolution (LTE) as the baseline technology to realize authentication and admission control, resource coordination, and interference management among different communication cells (generic functionalities). It operates in a larger coverage area (e.g., a factory hall) and handles functionalities requiring longer timescales. On the second tier, local radio coordinators operate in a smaller area and on a much more granular timescale, that is, enabling the required low-latency and high-reliability transmissions (mission-critical functionalities). Local coordinators can operate in two modes: a "centralized" mode, where both the user and control plane messages are transmitted via the local coordinators, and an "assisted device-to-device (D2D)" mode, where user and control planes are separated. The latter allows the direct exchange of user data between

Bernd Holfeld, Dennis Wieruch, Thomas Wirth, and Lars Thiele are with Fraunhofer Heinrich Hertz Institute; Shehzad Ali Ashraf, Jörg Huschke, Ismet Aktas, and Junaid Ansari are with Ericsson Research.

0163-6804/16/$25.00 © 2016 IEEE

devices (e.g., sensors and actuators) while control information routes via local coordinators. For critical factory applications, the assisted D2D mode may provide advantages over the centralized mode with respect to latency, thanks to the gains due to proximity of devices and a reduced number of communication hops.

In the following sections, we outline our major findings in relevance to the KoI project. We focus here on the air interface design of the local radio coordinator requiring reliable low-latency transmissions. We start with a discussion about the state-of-the-art technologies used in factory environments and their limitations in achieving the low-latency and high-reliability requirements. Furthermore, we describe the proposed medium access control (MAC) and PHY schemes for this specific use case in the context of 3GPP technologies. In particular, we discuss whether wireless communication for factory automation can be provided as an evolution of LTE; that is, it benefits from a backward-compatible mobile standard, or will require more substantial modifications toward a non-backward-compatible fifth generation (5G) system. Here, backward compatibility means that the legacy and 5G devices could share the same frequency carrier.

Requirements in Factory Automation

In the context of the KoI project, we have conducted a detailed questionnaire-based survey to gather first-hand information from notable industrial players involved in a broad range of factory automation processes. Table 1 summarizes the key findings from the survey in terms of the communication requirements. Depending on the specific application scenario, these requirements may differ within the shown ranges. To exemplify, we particularly provide the requirements of two factory applications being targeted by the wireless community in the table.

Although wireless communication offers several advantages over wired networks, it is not adequately leveraged in factory automation scenarios. Among others, this can be attributed to the fact that the currently available wireless technologies do not fulfill the ultra-high reliability requirements of $1 - 10^{-9}$ with very low (e2e) latency bounds down to 1 ms needed by factory automation applications. As a comparison, current cellular systems, such as LTE, are optimized for mobile broadband (MBB) traffic and target a BLER of 10^{-1} before retransmission with (e2e) latency bounds of several milliseconds. In addition, the factory automation use case is not only different due to the reliability and latency requirements, but also due to very different propagation conditions as discussed below, traffic characteristics (e.g., periodic and sporadic), and deployment peculiarities.

State-of-the-Art-Technologies and Their Limitations

Currently, factory automation applications are heavily dominated by wired technologies such as PROFIBUS/ PROFINET, SERCOS, HART, and CAN [3]. However, in wireless domain, the most commonly used factory communication solutions rely on customized radio stacks based on IEEE layer 1 (L1) and layer 2 (L2) technologies, as listed in Table 2. These IEEE-based standards operate in the unlicensed bands (below 6 GHz) and hence suffer from potential interference from other collocated networks sharing the same wireless spectrum.

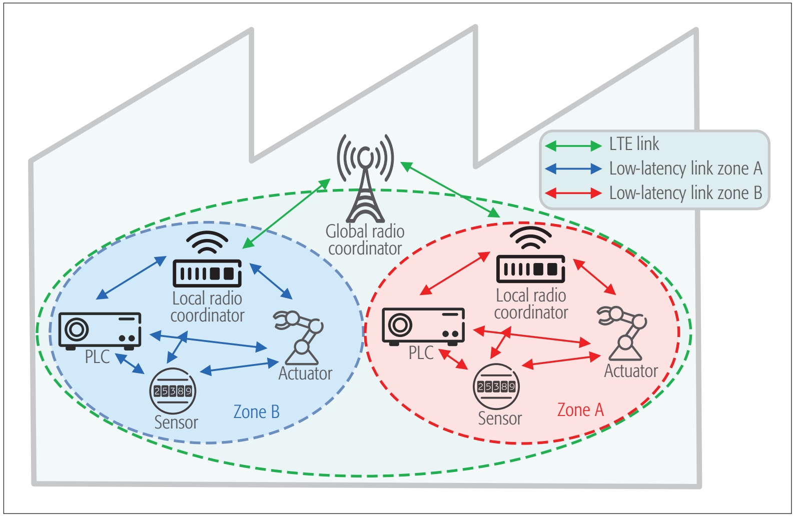


Figure 1. A communication architecture for wireless factory automation as envisioned in the project KoI [2].

Spectrum availability in unlicensed frequency spectrum inhibits guaranteed medium access and limits the scalability of the deployed solutions. This is associated with the regulatory policies for unlicensed spectrum that mandate features such as listen-before-talk (LBT), restriction of radio duty cycles, and transmit power limitation in order to facilitate coexistence. Under these policies, the stringent timing and reliability requirements of the C-MTC use case cannot be fulfilled. While wireless technologies operating in the licensed frequencies seem highly promising, these have still not surfaced for several C-MTC use cases, especially for factory automation applications. Operation in the licensed spectrum permits high transmit power levels and does not suffer from the drawbacks of mandatory coexistence regulations. Hence, it enables the implementation of deterministic medium access schemes. 3GPP's licensing-based LTE standard brings inherent advantages to fulfill the requirements of C-MTC applications. It allows deterministic multi-user scheduling by utilizing frequency multiplexing per time instance, support for quality of service, and flexible interference management for multiple cells. Therefore, by putting a focus on C-MTC besides the IoT-driven M-MTC market, 3GPP can deliver a favorable air interface for wireless factory automation and could certainly provide a single flexible solution for various requirements in this application area. In the following, we compare LTE to the currently used wireless standards for factory automation.

In Table 2, we highlight several layer 1/2 (L1/L2) features of the current LTE Release 12 and IEEE-based wireless technologies for factory automation. While LTE was initially designed for cellular macro networks with inter-site distances in the kilometer range, including support of up to 100 km coverage nowadays, other wireless technologies were specifically adjusted for

The physical layer and medium access mechanisms are major contributors to end-to-end latency from the radio communication perspective. Therefore, it is important to design lower layer protocols imparting as little latency as possible for mission-critical applications while fulfilling the reliability requirements.

e2e latency	Reliability	Data size	Communication range between devices	No. of devices per factory hall	Machine mobility (indoors)
Summarized results					
1 to 50 ms	$1 - 10^{-6}$ to $1 - 10^{-9}$	10 to 300 bytes	2 to 100 m	10 to 1000	0 to 10 m/s
Application scenario: Manufacturing processes					
< 10 ms	$1 - 10^{-9}$	< 50 bytes	< 100 m	< 1000	~ 1 m/s
Application scenario: Automated guided vehicles					
10 to 50 ms	$1 - 10^{-6}$ to $1 - 10^{-9}$	< 300 bytes	~ 2 m	< 1000	< 10 m/s

Table 1. Communication requirements for wireless factory automation gathered within the KoI project.

short- to medium-range communications below 200 m. Consequently, the symbol duration of IEEE-based technologies is much smaller, which in turn affects the (e2e) latency. However, to achieve a fair comparison of the latency impact between LTE and IEEE-based standards, we need to consider more than the pure time symbol duration. First, in IEEE-based standards the channel occupation time for a single transmitter is governed by additional symbols for synchronization preambles, control signaling, and LBT backoff, while LTE allows spreading data and control information over the frequency domain within the time symbol duration. Second, only LTE exploits multi-user access by frequency and space multiplexing, whereas IEEE-based standards predominantly rely on user multiplexing over time. Therefore, increasing the number of nodes has a significant impact on (e2e) latency for these standards. With increasing number of nodes, the LBT-based non-deterministic medium access schemes, for example, carrier sense medium access with collision avoidance (CSMA/CA), start to be inefficient. Even if the deterministic slotted medium access mode is used instead of CSMA/CA, the IEEE technologies can only support limited numbers of users to meet the latency requirements. In short, when comparing both technologies, we identify LTE as the choice for latency-critical factory applications.

Besides latency, factory automation also demands ultra-reliable transmission, as mentioned previously. Reliability is ensured by the use of forward error correction (FEC) schemes and exploiting the diversity gains. However, as specified in [9], LTE transmission is typically configured to operate for the target BLER of 10^{-1} before retransmission, which is considered to be a good trade-off between latency and system capacity for MBB services. In addition, it is to be noted that the turbo codes chosen for LTE have an error floor. For small packet size and code rate, the floor is below BLER of 10^{-5}. The coding efficiency decreases with the number of decoding iterations that can be executed within the tight latency requirement typical in C-MTC. Low-complexity convolutional codes that do not experience error floors and show similar performance for user data with small packet sizes are promising candidates. We study changes in coding and other L1/L2 modifications that are needed to enable mission-critical services using LTE later.

Key Design Features

The key design targets for C-MTC include low latency and ultra-high reliability. This requires exploiting certain design features at both layer 1 and layer 2 of the communication system, which are briefly described in the following. Please note that these design principles hold for C-MTC in general and factory automation in particular.

Enabling Ultra-Low Latency

For a communication system, delays at various protocol layers contribute to the (e2e) latency. The major contributors to the latency include protocol stacks, signal processing, medium access, transmission, and propagation delays. Processing delays are governed by the encoding and decoding complexity of the data at transmitter and receiver side, respectively. The physical layer and medium access mechanisms are major contributors to end-to-end latency from the radio communication perspective. Therefore, it is important to design lower layer protocols imparting as little latency as possible for mission-critical applications while fulfilling the reliability requirements.

Enabling Ultra-High Reliability

Diversity is one of the most significant techniques of the PHY layer for achieving highly reliable communication in a fading channel. As shown in [10], Fig. 2 illustrates that without diversity gains, a 90 dB margin is needed for guaranteeing lower than 10^{-9} probability of fading-induced outage. With diversity orders 8 and 16, the needed margin reduces to 18 dB and 9 dB, respectively. Time, frequency, and/or space are the three dimensions of achieving such high diversity gains. However, time diversity is not considered to be a suitable option for applications with strict latency bounds. It is also shown in [10] that hybrid automatic repeat request (HARQ) gains are not significant with such low latency requirements. Nevertheless, frequency diversity can be exploited on top of spatial diversity for C-MTC. Since D2D transmission cannot exploit high diversity gains only via spatial diversity, frequency diversity is of particular importance for the assisted D2D mode. Further details on diversity are described later.

L1/L2 technology	IEEE 802.11n (WLAN)	IEEE 802.11ac (WLAN)	IEEE 802.15.1 (WPAN)	Bluetooth 4.2 (WPAN)	IEEE 802.15.4 (WPAN)	3GPP LTE Rel-12 (4G cellular)
Industrial solution/standard	IWLAN		Bluetooth 1.2, WISA	Bluetooth	ZigBee, ISA100.11a, WirelessHART	
Spec. release	2009	2013	2005	2014	2011	2015
Range	< 200 m	< 200 m	< 100 m	< 100 m	< 10 m	< 100 km
Licensing	Unlicensed	Unlicensed	Unlicensed	Unlicensed	Unlicensed	Licensed
User multiplexing	Time	Time, space	Time	Time	Time	Time, space, frequency
Antenna support	4	8	1	1	1	8
Target error rate	PER: 0.1	PER: 0.1	BER: 0.001	BER: 0.0002 - 0.001	PER: 0.01	BLER: 0.1
FEC	Convolutional code, LDPC, STBC	Convolutional code, LDPC, STBC	No FEC, repetition-code, Hamming code	No FEC, repetition-code, Hamming code	No FEC, convolutional code, RSC	Turbo code, STBC
Frequency band	2.4 GHz, 5 GHz	5 GHz	2.4 GHz	2.4 GHz	780 MHz, 868 MHz, 915 MHz, 950 MHz, 2.45 GHz	400 MHz–4GHz
Bandwidth	20 MHz–40 MHz	20 MHz–160 MHz	1 MHz	1 MHz	200 kHz–5 MHz;	1.4 MHz–100 MHz
Time symbol duration	3.6 μs	3.6 μs	1 μs	1 μs	> 6 μs	71.4 μs
Theoretical peak data rate	< 600 Mb/s	< 6.93 Gb/s	1 Mb/s	24 Mb/s	< 1 Mb/s	DL: < 4 Gb/s UL: < 1.5 Gb/s
Signaling	OFDM	OFDM	Single carrier with spread spectrum	Single carrier with spread spectrum	Single carrier with spread spectrum	DL: OFDMA UL: SCFDMA
Modulation	Up to 64-QAM	Up to 256-QAM	GFSK	PSK, GFSK	Chirp-SK/FSK/PSK/ASK	DL: up to 256-QAM UL: up to 64-QAM
Channel access scheme	CSMA/CA and slotted	CSMA/CA and slotted	Slotted	Slotted	CSMA/CA and slotted	Scheduled

LDPC: Low-density parity check
STBC: Space-time block coding
RSC: Reed-Solomon code

Table 2. Layer 1 and 2 features of relevant wireless standards retrieved from standardization documents of IEEE 802.11n 2009 [4], IEEE 802.11ac 2013 [5], IEEE 802.15.1 2005 [6], IEEE 802.15.4 2011 [7], Bluetooth core v4.2 [8], and LTE 3GPP Rel-12 [9].

Radio Channels in Factory Environments

In factory automation, the communication system needs to be adapted for indoor radio propagation. Here, this means the building structure, that is, the existence of metallic ceilings and open metallic joists, as well as close-by production cells comprising active machine tools and industrial robots alter the scattering and reflection characteristics of the radio channel. Therefore, when proposing L1/L2 modifications for an LTE-based C-MTC air interface, the time dispersion of the multi-path channel and the evolution of the wireless signal over time are of high relevance. These parameters deliver design constraints on the minimum symbol length needed for the transmission without inter-symbol interference (ISI) and improved link adaptation.

Motivated by existing technologies in unlicensed spectrum, the use of spectrum below 6 GHz is the current choice for the automation industry. To gain insight into typical channel characteristics, we performed a wideband channel measurement campaign within the KoI project, recording the channel delay statistics in a representative factory

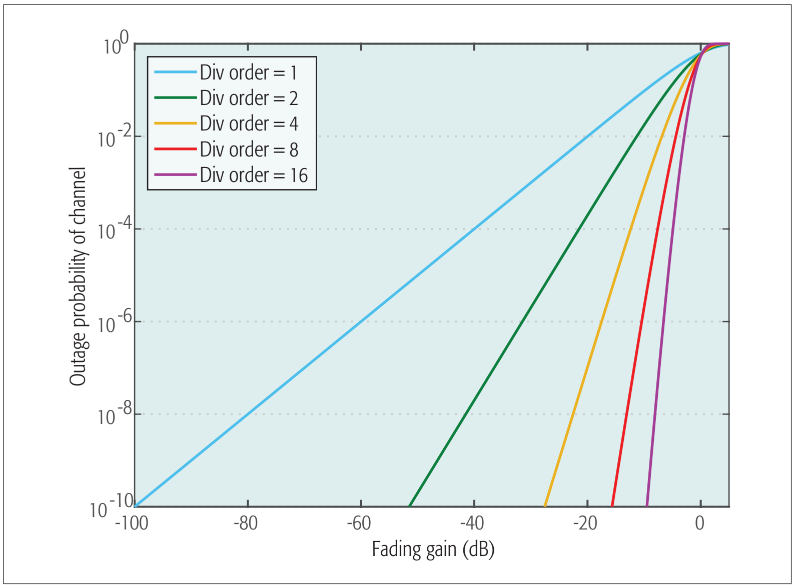


Figure 2. Outage probability of Rayleigh fading channels or different diversity orders [10].

automation cell at 5.8 GHz [11]. This carrier frequency was chosen due to its availability without licensing and complements other measurements performed for industrial applications at 2.4 GHz. Although the measurements were performed in unlicensed frequency bands of 5.8 GHz, the results could also be extended for neighboring frequency bands under a licensed regime. The addressed application scenario was the wireless control of industrial robots that repeatedly performed a pick-and-place process on a predefined trajectory at almost constant speed. The communication link was observed between a robot control unit placed within the automation cell and an antenna installed at the gripper of the moving robot arm. The 90th-percentile excess delay for non-line-of-sight (NLOS) transmission was 202 ns, but in a few cases of this setup the channel excess delays reached up to 350 ns (Fig. 3). To evolve LTE for communication in factory halls, the cyclic prefix (CP) length could be adapted based on these lower excess delays. Furthermore, the results also give a baseline for the D2D waveform design. Figure 3 exemplifies the power-delay profile of a measurement snapshot in which we estimated the set of dominant multipath components from the recorded channel impulse response (CIR). Also, we observed from the measurement data a high correlation between channel snapshots at the same positions of the repeated manufacturing process over time. The knowledge of the future position combined with the prediction of the link quality by previous channel estimates offers advantages in the design of feedback mechanisms and control data aspects. Channel quality indicator (CQI) feedback could be reduced while preserving high reliability and performance of the communication link. Hence, precise channel forecasts facilitate improved but simplified link adaptation in terms of pre-selection of modulation and coding schemes and optimized scheduling decisions. As a consequence, the L1/L2 processing of the air interface can be optimized for low-latency and reliable radio access for short-range industrial radios.

L1/L2 Modifications within LTE

Taking into account the aforementioned requirements, design principles, and channel peculiarities for the factory automation use case, this section outlines the key modifications required and/or being considered in LTE systems from the layer 1 and layer 2 perspectives.

Transmission Time Interval Shortening

The total latency between the time when data arrives in the transmission buffer and the time when a packet is delivered at the receiver is typically several times larger than 1 transmission time interval (TTI). For instance, in the case of uplink transmission, the device may also first send a scheduling request (SR) and wait for the uplink resource allocation from the base station. Hence, in order to achieve the requirement of 1 ms latency with an LTE system, the TTI should be redesigned to be significantly smaller.

LTE Release 13 is currently investigating the concept of TTI shortening for latency reduction. A TTI can be defined as the duration of an independent decodable transmission. Since in LTE systems the processing times are based on the TTI, a shorter TTI results in faster processing times. Therefore, TTI shortening leads to multiple benefits including lower transmission time, faster HARQ retransmissions, and lower processing time. Hence, in factory automation scenarios with typically small data size, TTI shortening would help in achieving lower latency. Moreover, TTI shortening allows scheduling flexibility as more user equipment (UE) can be scheduled in the same subframe using the same frequency resource. While current LTE systems use a TTI of 1 ms, LTE Release 13 is considering a TTI of duration 0.5 ms, and even the minimum possible duration consisting of only 1 OFDM symbol (i.e., 71.4 μs including the cyclic prefix). We believe that by shortening the TTI to 1 OFDM symbol, the minimum required e2e latency of 1 ms in factory automation can be fulfilled. Nevertheless, the cyclic prefix in LTE is optimized for macro scenarios operating at traditional LTE frequencies, which is not an efficient design regarding the factory deployments described above. We give further details on design complexities with respect to TTI shortening later.

Instant Uplink Access

Primarily, the LTE link layer is not designed to address latency-critical communication requirements. In an LTE system, the channel access and radio resource management are centrally coordinated by the LTE base station, or eNodeB. While the eNodeB is able to efficiently handle downlink transmissions, uplink transmissions involve high signaling overhead leading to undesired communication latency. For an uplink transmission, the device first needs to send a scheduling request (SR). Corresponding to the SR, the eNodeB sends the scheduling grant (SG) to the device, thereby indicating the schedule and the resources to be used. Finally, the user can transmit its data only after receiving the SG. Hence, the cycle of SR-SG-data induces a high degree of latency (at best this can be on average 9.5 ms with 1 ms TTI size). In order to deal with the uplink dynamic scheduling, the concept of so-called instant uplink access (IUA) is being investigated in LTE Release 13, where the SR-SG overhead is proposed to be eliminated. The eNodeB reserves prior uplink resources for a given device, and when the data arrives, it is directly sent out without any SR. While blocking some uplink resources in every subframe results in lower resource utilization when there is no uplink traffic, this scheme helps in substantially reducing the uplink latency. Besides latency reduction, system-level simulation results indicate that IUA also allows lower energy consumption for the device [12].

It is to be noted that the IUA concept is complementary to TTI shortening. These two concepts can lead to minimizing (e2e) latency of the existing LTE system, and thus are very well suited to the requirements of a wide range of factory automation applications.

Modulation and Coding Schemes

Modulation and coding selection impacts both the required received signal power and the required bandwidth [13]. In general, higher modulation order and code rate require additional signal power, but reduce the needed bandwidth. When it comes to modulation, there are practical limitations, such as transmitter and receiver impairments, which typically limit the highest modulation order. Considering LTE, the available modulation levels are quadrature phase shift keying (QPSK), 16-quadrature amplitude modulation (QAM), 64-QAM, and 256-QAM.

Although many modern communication systems such as LTE and IEEE 802.11ac use turbo or LDPC codes as FEC for data (Table 1), it is preferred to use convolutional codes in the low-latency and high-reliability C-MTC use cases such as factory automation. Convolutional codes have similar performance as turbo and LDPC codes for small block lengths that are typical for this use case (e.g., up to a few hundred bits). In contrast to convolutional codes, turbo and LDPC codes have an error floor, which makes these codes less efficient when the BLER reaches very low levels (e.g., 10^{-9}). Considering latency, decoding convolutional codes imparts shorter delay compared to the iterative decoders typically used for turbo and LDPC decoding. This is due to lower decoding complexity, and the property of convolutional codes that the decoder can process the code block while it is being received, thereby obtaining the decoded bits with very little delay. This requires that interleaving is only performed over frequency and not over time. However, for control channels that have block lengths smaller than 10 bits, block codes are preferred due to better performance and manageable decoding complexity [14].

Diversity

Another important use of coding is to harvest diversity. As discussed previously, diversity is a powerful tool for achieving high reliability, and to achieve spatial and frequency diversity in an OFDM system, it is essential to spread the coded bits over different diversity channels. Ideally, if the correct and erroneous code words differ in d positions, it is desired that these d positions are mapped to independent frequency bins or transmit antennas. If a deployment has M diversity channels, the code rate needs to be low enough to have free distance (convolutional codes) or minimum Hamming distance (block codes) sufficiently larger than M.

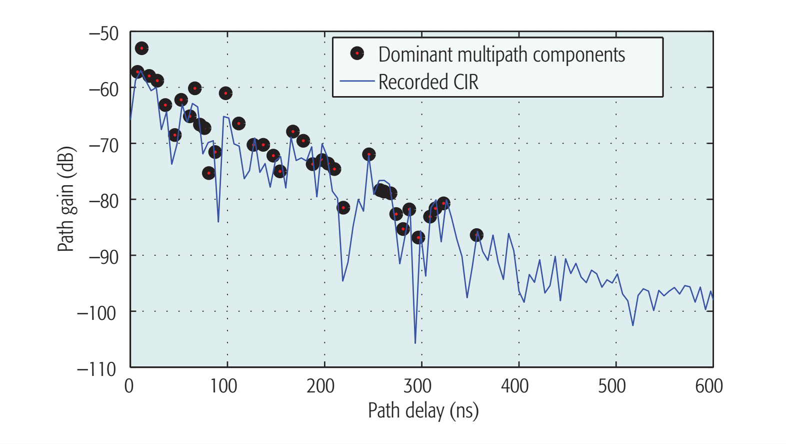


Figure 3. Typical power-delay profile obtained during the KoI channel measurement campaign in a factory automation setup at 5.8GHz [11].

In order to enable multi-cell factory deployments to work with a small frequency reuse factor, not only is the SNR of importance, but the system also has to work reliably at a low average SINR. A BLER of 10^{-9} at low average SINR of 3–10 dB can only be achieved with a very high diversity order of 16 (e.g., 8 × 2 or 2 × 8 antennas), unless transmitter-side channel state information is available. This can be seen in Fig. 4, assuming for simplicity that the SNR and SINR requirements for the same BLER are equal. We consider here up to 8 transmit antennas and a code rate of 1/2 with a free distance of 10 combined with Alamouti code, which can exploit the transmit-side diversity order to a large extent. However, there is a diversity loss associated with the use of Alamouti code for more than two transmit antennas. For example, with BLER of 10^{-9}, the required SNR for 8 transmit antennas (using binary PSK, BPSK, and a code rate of 1/2) is approximately 2.3 dB higher compared to when maximum diversity gains can be exploited, (i.e., ideal full diversity).

An 8 × 2 configuration is possible for communication where one endpoint is a base station that typically has the deployment space and complexity constraints allowing for a high number of antennas. For D2D, where only a few antennas can be accommodated, such high diversity order might not be achieved even after exploiting gains from frequency diversity.

Performance degradation due to antenna correlation is also shown in Fig. 4. We illustrate the effect of receive antenna correlation for an 8 × 2 antenna configuration as well as the effect of transmit antenna correlation for a 2 × 8 antenna configuration. We see that with antenna correlation of 0.5, the BLER slope is approximately the same as when there is no antenna correlation, although there is a small SNR penalty at lower BLER. This indicates that the diversity order is very well preserved. With antenna correlation

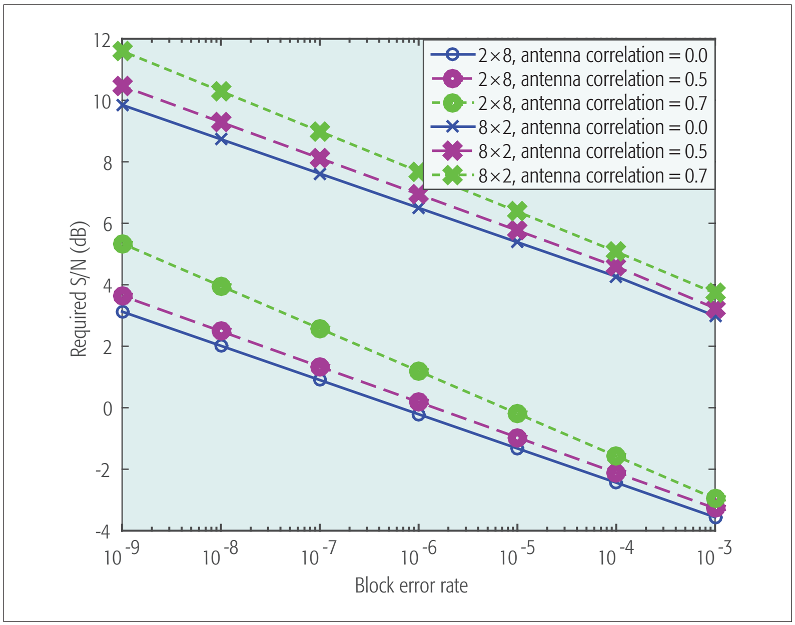


Figure 4. Impact of antenna correlation on link performance for the 8 × 2 and 2 × 8 antenna configurations with QPSK and rate 1/2 coding [10].

as high as 0.7, there is a more noticeable effect on the BLER slope. However, even in this case, the SNR penalty at 10^{-9} BLER is approximately 2 dB.

5G Outlook: Beyond the LTE Evolution

For factory automation, we target the TTI to be on the order of 100 μs to satisfy the most stringent use cases. If the LTE TTI is reduced to 1 OFDM symbol of 71.4 μs, this target could be met. However, having just a single OFDM symbol per TTI comes with the drawback of a small number of resource elements. Given that the amount of control signaling and reference symbols has to stay approximately constant regardless of the TTI duration, the overhead percentage increases accordingly. Furthermore, in factory environments, the delay spread and excess delay are much lower than those for which the LTE OFDM symbol duration was originally designed. Therefore, a scaled version of LTE numerology is more suitable for latency-critical systems in the factory environment (e.g., using a scaling factor of 5). Thus, the new scaled numerology being considered for 5G has a subcarrier spacing of 75 kHz and OFDM symbol duration of 13.3 μs excluding the cyclic prefix (CP). Moreover, early decoding is considered to be an important aspect for delay sensitive communication. Hence, the control signaling and reference symbols need to be placed at the start of the subframe to allow early decoding of the received data [14].

Although orthogonal waveforms such as LTE's OFDM have many advantages, some limitations could be addressed by a novel waveform in 5G [15]. OFDM is well suited for cellular systems. However, for D2D communication, it suffers from the stringent synchronization requirements. For cellular systems, multiple access interference within a cell is avoided in the downlink by inherently synchronized transmission. In the uplink, synchronized reception of the signals from multiple devices is achieved by adjusting the timing advance for each device accordingly. For D2D communication, if multiple D2D links are concurrently active and in interference range of each other, the timing advance of multiple transmitting devices can in general not be controlled in such a way that synchronous reception at all receiving devices is achieved.

Here, D2D systems could benefit from multi-carrier waveforms with improved frequency localization and relaxed synchronization requirements. Filtering the transmit signal achieves lower spectral side-lobes. In order to deal with the interference between asynchronous receptions on adjacent sub-bands that emerges in OFDM, the filtering could be done separately on each transmitted sub-band, using filtered OFDM. Alternatively, filtering each resource block separately as in universal filtered multicarrier (UFMC) has the benefit of not requiring allocation-dependent filters. Furthermore, filter bank multi-carrier (FBMC) could be used. This generalization of multi-carrier modulation introduces a well designed poly-phase prototype filter shape onto the modulated signal on each subcarrier. FBMC systems benefit from cyclic-prefix-free transmission, saving additional resources. Nevertheless, the drawback of the FBMC approach is degraded time-localization behavior. In particular, multiple symbols overlap in the time domain and increase the effective symbol duration. Therefore, it first needs to be analyzed if the FBMC waveform can meet the low-latency requirements of the indoor factory automation applications with short channel impulse responses. In conclusion, going with a filtered OFDM design is currently the best and most mature choice for 5G factory communication systems because of the time localization specifics of FBMC.

Summary

This article discusses the use of wireless communication in a factory automation scenario and presents the challenging communication requirements. While current wireless solutions for such applications are dominated by proprietary implementations and are mostly applied in isolated scenarios, a worldwide wireless standard could leverage the advantages of going wireless in a global market. Licensed operation naturally brings advantages in meeting the latency and reliability requirements by excluding the need to handle coexisting systems. By fulfilling the requirements of mission-critical applications, wireless technologies based on LTE will certainly enable new services in factory automation. For this, the current LTE design needs to undergo some modifications, which are partially being considered in 3GPP already and discussed in this article. While the proposed design modifications could address a broad range of factory automation services, a few mission-critical applications require revolutionary (non-backward-compatible) changes in the communication system as highlighted above. Either way, whether it be LTE's evolution or 5G that makes it to the market of factory automation, in both cases it opens up new business opportunities for vendors and operators using 3GPP technologies.

Acknowledgment

This work is part of the research project Koordinierte Industriekommunikation (KoI) under grant support of the Federal Ministry of Education and Research (BMBF) of Germany. Partners involved in the project are Ericsson, Fraunhofer Heinrich Hertz Institute, RWTH Aachen University, the University of Paderborn, WiseSense GmbH, and WeissRobotics GmbH & Co. KG.

References

[1] E. Dahlmann *et al.*, "5G Wireless Access: Requirements and Realization," *IEEE Commun. Mag.*, vol. 52, no. 12, Dec. 2014, pp. 42–47.

[2] Research project Koordinierte Industriekommunikation (KoI) supported by the Federal Ministry of Education and Research (BMBF) of Germany, Duration: 01/2015–06/2017; http://www.koi-projekt.de

[3] J.-P. Thomesse, "Fieldbus Technology in Industrial Automation," *Proc. IEEE*, vol. 93, no. 6, June 2005, pp. 1073–1101.

[4] IEEE Std 802.11n, "Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications; Amendment 5: Enhancements for Higher Throughput," Sept. 2009; http://standards.ieee.org

[5] IEEE Std 802.11ac, "Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications; Amendment 4: Enhancements for Very High Throughput for Operation in Bands below 6 GHz," Dec. 2013; http://standards.ieee.org.

[6] IEEE Std 802.15.1, "Part 15.1: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Wireless Personal Area Networks (WPANs)," June 2005; http://standards.ieee.org.

[7] IEEE Std 802.15.4, "Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs)," Sept. 2011; http://standards.ieee.org.

[8] Bluetooth Special Interest Group, "Covered Core Package Version: 4.2," Bluetooth System Spec., Dec. 2014.

[9] 3GPP TS 36.201 (v12.2.0), "Evolved Universal Terrestrial Radio Access (E-UTRA); LTE Physical Layer; General Description (Release 12)," Mar. 2015.

[10] N. A. Johansson *et al.*, "Radio Access for Ultra-reliable and Low-Latency 5G Communications," *Proc. IEEE ICC Wksp. 5G & Beyond – Enabling Technologies and Applications*, London, U.K., June 2015.

[11] B. Holfeld *et al.*, "Radio Channel Characterization at 5.85 GHz for Wireless M2M Communication of Industrial Robots," *Proc. IEEE Wireless Commun. and Networking Conf.*, Doha, Qatar, Apr. 2016.

[12] 3GPP TR 36.881 (v0.5.0), "Evolved Universal Terrestrial Radio Access (E-UTRA); Study on Latency Reduction Techniques for LTE (Release 13)," Nov. 2015.

[13] O. N. C. Yilmaz *et al.*, "Analysis of Ultra-Reliable and Low-Latency 5G Communication for a Factory Automation Use Case," *Proc. IEEE ICC Wksp. 5G & Beyond Enabling Technologies and Applications*, London, U.K., June 2015.

[14] S. A. Ashraf *et al.*, "Control Channel Design Trade-offs for Ultra-Reliable and Low-Latency Communication System," *Proc. IEEE GLOBECOM Wksp. Ultra-Low Latency and Ultra-High Reliability in Wireless Communication*, San Diego, CA, Dec. 2015.

[15] G. Wunder *et al.*, "5GNOW: Non-Orthogonal, Asynchronous Waveforms for Future Mobile Applications," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 97–105.

Biographies

Bernd Holfeld (bernd.holfeld@hhi.fraunhofer.de) joined the Fraunhofer Heinrich Hertz Institute, Berlin, Germany, in 2012 after completing his Dipl.-Ing. (M.Sc.) degree in electrical engineering at Technische Universität Dresden, Germany, the same year. In 2010, he was a visiting student researcher at the Institute for Telecommunications Research at UniSA, Adelaide, Australia. He has been involved in several research projects on 4G and 5G radio access. His current focus is on machine-to-machine communication for industrial wireless systems.

Dennis Wieruch (dennis.wieruch@hhi.fraunhofer.de) received his Dipl.-Ing. (M.Sc.) degree in computer engineering from Technische Universität Berlin, Germany, in 2009. In March 2009, he joined the Fraunhofer Heinrich Hertz Institute, Germany, as a research associate, where he is currently pursuing a Ph.D. degree. His research interests include wireless communications with a focus on cognitive radio, compressed sensing, software defined radio, and digital signal processing. He has participated in several European Union and German government funded research projects.

Thomas Wirth (thomas.wirth@hhi.fraunhofer.de) received his Dipl.-Inform. (M.Sc.) degree in computer science from the Universität Würzburg, Germany, in 2004. He joined the Fraunhofer Heinrich Hertz Institute in 2006, working as a senior researcher on LTE and LTE-Advanced technologies. Since 2011, he has headed the Software Defined Radio group with a focus on PHY and MAC algorithms and real-time implementation for future wireless networks, including industrial radio applications and 5G prototypes.

Lars Thiele (lars.thiele@hhi.fraunhofer.de) is currently associated with the Fraunhofer Heinrich Hertz Institute, where he leads the System Level Innovation group. He received his Dipl.-Ing. (M.Sc.) degree in electrical engineering from Technische Universität Berlin in 2005 and his Dr.-Ing. (Ph.D.) degree from Technische Universität München in 2013. He has co-/authored more than 50 papers and several book chapters in the fields of radio propagation modeling, large-scale system-level simulations, CoMP transmission, and massive MIMO.

Shehzad Ali Ashraf (shehzad.ali.ashraf@ericsson.com) is an experienced researcher at Ericsson Research, which he joined in 2013. He holds an M.Sc. in electrical engineering from RWTH Aachen University. Since joining Ericsson, he has been deeply involved in European Union and German government funded research projects related to the development of 5G concepts. Currently, he is also involved in 3GPP standardization. His research interests include 4G and 5G radio access technologies, and machine-type communications.

Jörg Huschke (joerg.huschke@ericsson.com) is a master researcher in radio research at Ericsson Research. He has been working toward 3GPP standardization of MBMS for LTE and its performance evaluation. In 2016, he began leading the radio interface team in the German research project KoI on evolved LTE and 5G-based wireless communication in factory automation. He received his diploma degree in electrical engineering from Georg-Simon-Ohm University of Applied Sciences, Nuremberg, Germany, in 1993.

Ismet Aktas (ismet.aktas@ericsson.com) works at Ericsson Research on concept development and evaluation for 5G. In this context, he is looking into critical machine type communication in general and wireless factory automation in particular. He obtained his Ph.D. and Diploma degrees from RWTH Aachen University. Currently, he is leading the KoI project. His research interests include cross-layer design, jamming detection, machine type communication, and industrial automation.

Junaid Ansari (junaid.ansari@ericsson.com) is currently associated with Ericsson Research. Previously, he worked as a postdoctoral researcher and research assistant at the Institute for Networked Systems at RWTH Aachen University. He received his Ph.D. (2012) and M.Sc. (2006) degrees from RWTH Aachen University. He has worked on several collaborative European Union and German government funded research projects. His research interests include cognitive radios, embedded intelligence, and system architecture design for next generation wireless networks.

By fulfilling the requirements of mission-critical applications, wireless technologies based on LTE will certainly enable new services in factory automation. For this, the current LTE system needs to undergo some design modifications, which are partially being considered in 3GPP already and discussed in this article.

LTE Release 14 Outlook

Christian Hoymann, David Astely, Magnus Stattin, Gustav Wikström, Jung-Fu (Thomas) Cheng, Andreas Höglund, Mattias Frenne, Ricardo Blasco, Joerg Huschke, and Fredrik Gunnarsson

The authors provide an overview of foreseen key technology areas and components for LTE Release 14, including latency reductions, enhancements for machine-type communication, operation in unlicensed spectrum, massive multi-antenna systems, broadcasting, positioning, and support for intelligent transportation systems.

Abstract

Today's 4G LTE systems bring unprecedented mobile broadband performance to over a billion of users across the globe. Recently, work on a 5G mobile communication system has begun, and next to a new 5G air interface, LTE will be an essential component. The evolution of LTE will therefore strive to meet 5G requirements and to address 5G use cases. In this article, we provide an overview of foreseen key technology areas and components for LTE Release 14, including latency reductions, enhancements for machine-type communication, operation in unlicensed spectrum, massive multi-antenna systems, broadcasting, positioning, and support for intelligent transportation systems.

Introduction

The first release of the Long Term Evolution (LTE) specifications, Release 8, was completed in 2008, and commercial network operation already began in December 2009. This was soon followed by deployments on a global scale, and since then there has been unprecedented adoption worldwide. In the first six years of commercial availability, more than 440 LTE networks have been launched in over 145 countries, and at the end of 2015, the number of LTE subscriptions reached one billion [1].

Ever since the first release, the LTE specifications have been regularly updated in Releases 9 through 13, introducing enhancements and new features to improve efficiency, and to boost both user and system performance. This includes carrier aggregation for spectrum flexibility, advanced antenna techniques and advanced receivers to increase spectral efficiency, small cell enhancements to address densification, as well as WiFi interworking and licensed-assisted access to exploit unlicensed spectrum. At the same time, support for voice calls, public warning systems, positioning, and multimedia broadcast multicast services have been added in addition to enhancements for the Internet of Things (IoT) with optimizations for machine-type communications and support for public safety with device-to-device communication.

In short, LTE is continuously evolving, addressing not only mobile broadband, but also new areas and use cases. In light of the advancements, Release 13 and onwards is also referred to as "LTE-Advanced Pro." The work on Release 14 has started with the target of being finalized in March 2017.

At the same time, the work on the future fifth generation (5G) radio access is ongoing in industry and academia as well as in fora such as the International Telecommunication Union (ITU) and Third Generation Partnership Project (3GPP) [2–4]. Since LTE is an essential part of the future radio access, 3GPP has decided to continue strong evolution of LTE in parallel with the development of a new radio interface and also to submit the two to IMT-2020. Hence, the evolution of LTE in Release 14 and beyond will strive to meet and address corresponding 5G requirements and use cases, respectively.

In the present article, we review the requirements for future 5G radio access to identify some main technology areas and components for LTE evolution in Release 14 and beyond. These include:
- Latency reduction
- Enhanced operation in unlicensed spectrum
- Machine-type communication
- Massive multi-antenna systems
- Intelligent transportation systems (ITS)
- Enhanced multimedia broadcasting
- Enhanced positioning

The different areas are then discussed, followed by a conclusion in the final section. Note that the above list is not exclusive and that additional areas may become part of LTE Release 14 as work progresses.

Requirements for Future Radio Access

There is an industry consensus that the most important use cases for radio access in 2020 and beyond can be categorized in three families [3], which is also reflected in the ongoing 3GPP work on 5G requirements [4]:
- Enhanced mobile broadband (eMBB)
- Massive machine-type communication (mMTC)
- Ultra-reliable and low-latency communication (URLLC)

eMBB will require massive system capacity to meet the predicted future traffic growth. At the same time, future systems will not only offer higher peak rates up to 20 Gb/s, but more importantly offer much higher data rates in real-life deployments, for example, 10 Mb/s everywhere, several 100 Mb/s in dense urban environments, and even higher in hotspot environments. Densification with more network nodes, use of more

The authors are with Ericsson Research.

0163-6804/16/$25.00 © 2016 IEEE

spectrum, both licensed and unlicensed, as well as spectral efficiency enhancements are needed.

This motivates further evolution of licensed-assisted access (LAA) and massive multiple antenna systems as described below. With eight-layer multiple-input multiple-output (MIMO) transmission defined in Release 10 and carrier aggregation of up to 32 carriers introduced in Release 13, the LTE peak data rate can already go up to ~25 Gb/s. However, very high data rates will also call for latency reductions due to the properties of Internet protocols as outlined below. Furthermore, media content streaming constitutes a major part of the future traffic volumes, and this motivates further evolution of eMBMS as described below.

mMTC addresses applications with a very large number of sensors, actuators, and similar devices typically associated with little traffic as well as requirements on low device cost and very long battery life. Together with network enhancements, such as improved coverage and signaling reductions, they enable the vision of a networked society with billions of connected things. Challenges in terms of coverage, device cost, and battery life have been addressed in previous standard releases as described below, where further evolution is also outlined.

URLLC implies fulfilling very tough requirements on reliability, availability, and latency in order to offer connectivity that is essentially always available. Examples include health applications, traffic safety and control, control of critical infrastructures, and connectivity for industrial processes. More specifically, in Release 14, an intelligent transportation system based on LTE will be developed as described below. We also note that latency reduction will enable even more low-latency applications.

Furthermore, positioning enhancements will not only add value to all the above mentioned use cases but also be needed for emergency calls. We outline enhancements for positioning below.

Already today, LTE offers support for many diverse use cases, and from the above it is clear that LTE capabilities will be significantly extended, allowing LTE to address even more challenging use cases in the future.

Latency Reduction

While much attention has been paid to improving LTE data rates, little effort has been spent on reducing packet latency. However, the perceived throughput will be affected by both aspects; especially for smaller packets does the impact of latency become visible. For TCP traffic, reduced latency is essential since, particularly during the TCP slow start phase, the rate of TCP acknowledgments determines the achievable data rate. Furthermore, low latency is of key importance to enable URLLC use cases with tight delay requirements. Latency reduction techniques are currently being considered in 3GPP in the form of an extension of semi-persistent scheduling for faster uplink access, reduction of handover interruption time, as well as shorter transmission intervals and processing times [5]. For time-division duplexing (TDD) operation, latency can be further reduced by increasing the number of uplink-downlink (UL-DL) switches, which currently is at most one per 5 ms.

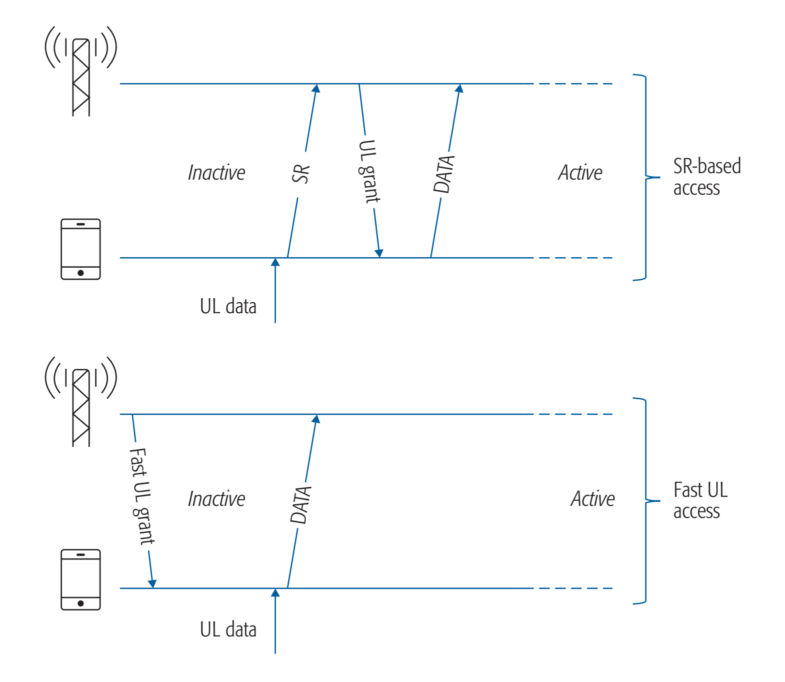


Figure 1. Conventional scheduling-request-based access (top) and access with a Fast UL grant (bottom).

In the UL, the medium access is based on scheduling requests (SRs) that are sent by the terminal to request resources if data needs to be sent. This introduces delay since the terminal must wait for an SR opportunity as well as the extra round-trip time needed to grant the transmission. By instead configuring a terminal with a periodic UL grant (e.g., with a 1 ms periodicity), referred to as a Fast UL grant [5], the terminal is allowed to transmit without the SR related delay. The existing SR-based UL access as well as UL access using a Fast UL grant are depicted in Fig. 1. To avoid unnecessary battery consumption and interference, the terminal should only use the UL resources if it has data to send. Additionally, to avoid resources being underutilized, they could be overbooked and assigned to multiple terminals with different reference signal settings to allow the network to distinguish them.

On the physical layer, the introduction of a shorter transmission time interval (TTI) reduces the data transmission and processing delays. It is essential, though, to reduce the period of control signaling in both UL and DL for scheduling commands and hybrid automatic repeat request (HARQ) feedback, and also consider reference signals for demodulation. TTI durations down to a single orthogonal frequency-division multiplexing (OFDM) symbol are being considered. However, since the control and reference signal overhead grows with decreasing TTI, the final design needs to find a good trade-off. It should also be noted that the solution is backward compatible in the sense that legacy terminals can be served on the same carrier.

The above enhancements have the potential to significantly reduce user plane latency. With Fast UL, the latency for a sporadic UL transmission can be reduced from ~12.5 ms down to ~7.5 ms for the current 1 ms TTI duration.

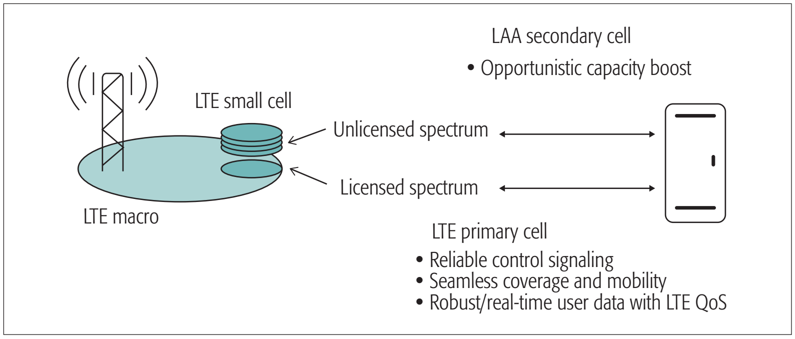

LAA secondary cell
• Opportunistic capacity boost
LTE small cell
Unlicensed spectrum
Licensed spectrum
LTE macro
LTE primary cell
• Reliable control signaling
• Seamless coverage and mobility
• Robust/real-time user data with LTE QoS

Figure 2. LAA enabling transmissions on secondary cell(s) operated in unlicensed spectrum controlled from a primary cell operating in licensed spectrum using carrier aggregation.

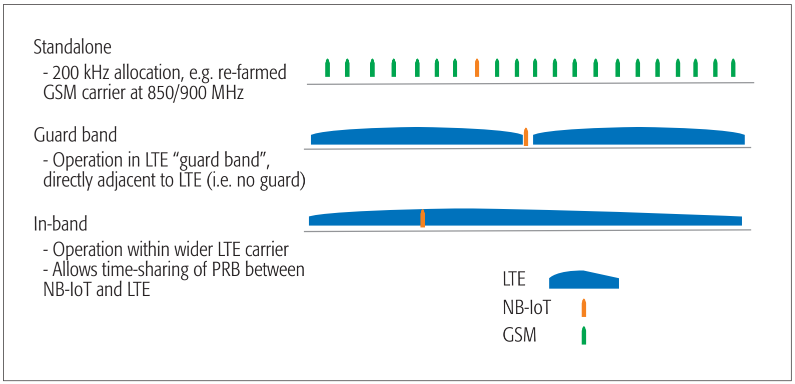

Standalone
- 200 kHz allocation, e.g. re-farmed GSM carrier at 850/900 MHz
Guard band
- Operation in LTE "guard band", directly adjacent to LTE (i.e. no guard)
In-band
- Operation within wider LTE carrier
- Allows time-sharing of PRB between NB-IoT and LTE
LTE
NB-IoT
GSM

Figure 3. NB-IoT deployment modes.

With a reduction of the TTI length down to, for example, 2 OFDM symbols, the one-way latency reduces to ~1 ms.

Licensed-Assisted Access

Existing and new licensed spectrum will remain fundamental for providing seamless wide-area coverage, achieving the highest spectral efficiency, and ensuring reliability of cellular networks. To meet the ever increasing data traffic demand, more spectrum will be needed. Given the large amount of spectrum available in the unlicensed bands, it is therefore of interest to use it as a complement to licensed spectrum. After careful study [6], 3GPP introduced LAA in Release 13 to enable LTE DL transmissions in secondary cells operated in unlicensed spectrum. These transmissions are controlled and coordinated from primary cells operating in licensed spectrum using the carrier aggregation framework (Fig. 2). This approach enables operators to enhance the seamless coverage in the LTE network with additional bandwidth and capacity. To achieve coexistence with other technologies operating in the same band, such as IEEE 802.11, listen-before-talk (LBT) procedures and discontinuous transmission with limited maximum channel occupancy time were introduced.

To further enhance the capabilities of LAA operations, UL channel access will be added in Release 14. Due to LBT procedures, several protocol enhancements may be introduced to improve LAA UL operation efficiency. Enabling one DL subframe to send grants for several UL subframes can reduce the overhead and significantly increase the throughput of LAA UL transmissions. Due to the uncertainty in channel access opportunities on carriers in unlicensed spectrum, it is more efficient for LAA UL HARQ to follow an asynchronous protocol, similar to LAA DL HARQ. The UL retransmission can be scheduled by a UL grant and occur at any time relative to the initial transmission.

On the physical layer side, 256-quadrature amplitude modulation (QAM) support will be introduced to bring LTE UL capability on par with other small cell technologies. Another potential enhancement to consider is the dual connectivity framework wherein the terminal may also simultaneously receive and transmit to a master and a secondary base station when the two base stations are connected via non-ideal backhaul. The combination of dual connectivity and LAA will extend unlicensed band LTE operations to even more deployment scenarios.

LTE-WLAN aggregation, introduced in Release 13, is another area where protocols will be enhanced in Release 14. Release 13 supports LTE-WLAN aggregation for the DL. Release 14 will allow aggregation for the UL as well. Additional information collection and feedback (e.g., better estimation of available WLAN capacity) as well as automatic neighbor relation procedures are to be introduced to improve performance.

Due to the discontinuous transmission with limited maximum channel occupancy time, LAA carriers can support very dynamic muting of otherwise persistent signals, such as cell-specific reference signals. Such lean operations can lead to reduced inter-cell interference and enhanced energy efficiency. It is therefore noted that such designs and benefits can also be extended to carriers in the licensed spectrum with significant user throughput improvements, especially in combination with higher order modulation such as 256-QAM.

Machine-Type Communication and the Internet of Things

Significant enhancements to LTE have been introduced to efficiently address the mMTC use case, that is, to allow very simple devices to be connected in a power-efficient manner. In Release 13, the existing track of MTC improvements [7] has been further evolved by reducing the device bandwidth to 1.4 MHz and the output power to 20 dBm, achieving even lower device cost. Furthermore, coverage enhancements up to 15–20 dB were introduced. These devices will still operate in all existing LTE system bandwidths.

Also in Release 13, another LTE-based track called narrowband IoT (NB-IoT) was standardized, with similar targets when it comes to coverage, power consumption, and device complexity. The smaller device bandwidth of 200 kHz reduces data rates and increases latency, but also offers greater deployment flexibility, as shown in Fig. 3. Such devices can be supported using one resource block inside an LTE carrier, in a standalone system with 200 kHz bandwidth, for example, on a re-farmed GSM carrier, or alternatively in the guard band of another (LTE) system.

The evolution of LTE will include enhancements common to both the MTC and NB-IoT solutions. More specifically:

- Point-to-multipoint transmissions enable, for example, efficient software / firmware updates or addressing of many actuators (e.g., light switches) for thousands of devices simultaneously.
- Positioning brings benefits to many different mMTC use cases. There is already support for positioning in Release13, but improving the positioning accuracy is desirable by introducing dedicated signals, procedures, and requirements for enhanced positioning performance in Release 14.
- Cost-efficient and easily deployable relays can reduce the transmission times, especially in challenging coverage conditions, and improve battery life, increase capacity, reduce latency, and enhance coverage simultaneously. Besides connecting sensor-type devices, relaying is also interesting for smart wearables and is considered in Release 14.

Furthermore, higher-layer enhancements for signaling reductions and extended DRX cycles introduced in Release 13 improve battery life for devices and improve network capacity. Longer DRX cycles in connected mode make it possible to keep terminals with high requirements on DL reachability in connected mode, but also put new requirements on, for example, mobility handling and congestion control in connected state. Work in Release 14 is foreseen to address these aspects, and to further reduce the radio and network interfaces signaling overhead, further extend battery life, and reduce the access latency for all types of devices.

Another potential Release 14 addition is NB-IoT operation in unpaired (TDD) spectrum, currently supported for the MTC track but not for NB-IoT.

As a final remark, due to the deployment flexibility, NB-IoT may be well suited for the migration into 5G, complementing the new radio access with the ability to support massive amounts of low-cost devices.

Massive Multi-Antenna Systems

In Release 13, a study of MIMO enhancements to extend the current support of 8 up to 64 transmit antennas was conducted, thus targeting a massive number of controllable antenna elements at the base station [8]. The study considered simultaneous horizontal and vertical adaptive transmission, utilizing two-dimensional antenna arrays with both closed loop and open loop (beamforming) operation modes. Significant performance improvement for both single-user and multi-user MIMO was found, which led to standard enhancements for up to 16 antennas in Release 13. This included feedback and small cell sounding enhancements for both closed and open loop operation in addition to an extension of the number of co-scheduled terminals to 8 for multi-user MIMO.

The closed loop mode is suitable for both FDD and TDD. It uses measurement channel state information reference signals (CSI-RS) per antenna and a precoder matrix codebook for terminal feedback (Fig. 4). A large number of terminals can be served in a cell without increasing the CSI-RS overhead as the CSI-RSs are cell-specific. However, for an increased number of antennas, the increasing CSI-RS overhead will at some point neutralize any possible massive MIMO gain. Open loop beamforming, where the beamforming direction is determined without explicit feedback from the terminal, has increased efficiency for large numbers of transmit antennas, since the CSI-RS are terminal-specific and beamformed (Fig. 5). However, when the number of served terminals becomes large, the overhead is problematic. Hence, the closed loop and open loop modes are complementary, and further enhancements to both of them will increase coverage and capacity further.

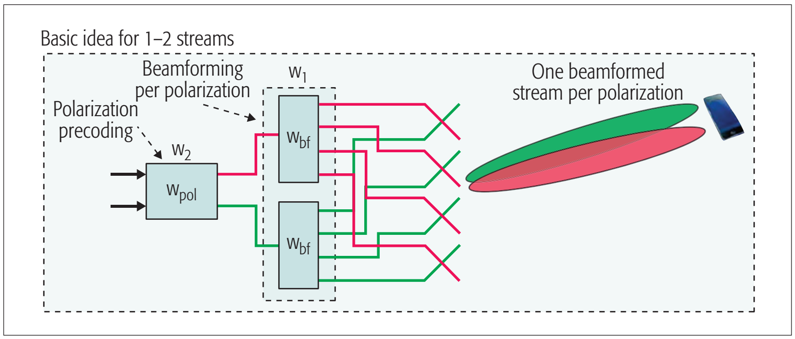

Basic idea for 1–2 streams
Beamforming per polarization
w_1
One beamformed stream per polarization
Polarization precoding
w_2
w_{pol}
w_{bf}
w_{bf}

Figure 4. Closed loop MIMO where the terminal controls the precoders W1 and W2 for robust operation.

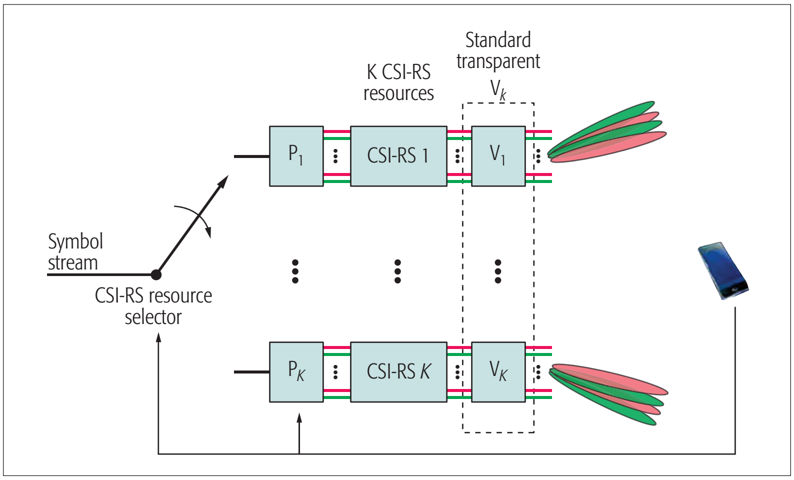


Figure 5. Open loop beamforming, where the terminal selects one out of K beams formed by the standard transparent precoders Vk, which are determined without explicit terminal feedback.

In Release 14, the MIMO evolution will continue, targeting up to 32 transmit antennas. Currently, both single- and multi-user MIMO performance is seen to be limited by the quality of the channel knowledge at the transmitter, and this motivates investigating new feedback methodologies for high-resolution feedback in addition to the existing precoding codebook-based scheme. The challenge is how to ensure sufficiently good transmitter channel knowledge for frequency-division duplexing (FDD) and TDD when full reciprocity is not available, for instance when the terminal has fewer transmit antennas than receive antennas.

For open loop operation, further beamforming enhancements are envisioned, introducing dynamic CSI-RS allocation, allowing for effi-

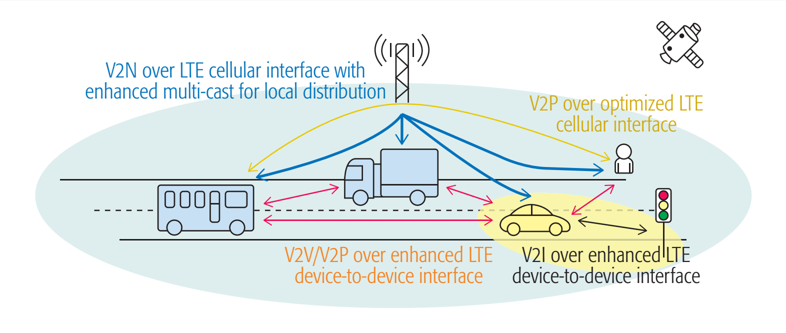

V2N over LTE cellular interface with
enhanced multi-cast for local distribution
V2P over optimized LTE
cellular interface
V2V/V2P over enhanced LTE
device-to-device interface
V2I over enhanced LTE
device-to-device interface

Figure 6. Illustration of the different ITS connectivity scenarios and the different LTE interfaces.

cient pooling of RS resources. This would then both manage the reference signal overhead with beamforming with a larger number of simultaneous terminals as well as provide robustness for the open loop mode.

Finally, studies on further enhanced DL coordinated multipoint operation, including beamforming coordination between multiple base stations using massive MIMO, are also foreseen.

Intelligent Transportation Systems

Radio communications are instrumental in enabling the deployment of ITS, which have been identified as a way of improving traffic safety and efficiency. To support these as well as many other applications, 3GPP is developing an ITS solution based on LTE targeting different vehicle-to-anything (V2X) connectivity scenarios, including vehicle-to-vehicle (V2V), vehicle-to-roadside infrastructure (V2I), vehicle-to-pedestrian (V2P), and vehicle-to-network (V2N) (Fig. 6). LTE V2X intends to reuse the higher layers and services specified by the European Telecommunications Standards Institute (ETSI), and hence specify only the lower layers. An LTE solution will benefit from the existence of an already deployed network infrastructure to support many of the use cases and provide an increased level of security over distributed systems.

The standardization started in Release 13 by identifying the uses cases of V2X services along with their requirements. This is followed by enhancing the existing interfaces that are necessary to support the connectivity scenarios [9].

The direct device-to-device interface used, for example, in V2V and V2P needs to support increased terminal mobility. At typical vehicle speeds, the transmitted signals are significantly degraded, especially at high carrier frequencies where, for example, the coherence time of the channel is much shorter than in traditional cellular communications. Thus, it is necessary to introduce a new LTE subframe structure with increased pilot-symbol density that allows for accurate channel estimation. Similarly, the cellular methods for obtaining frequency synchronization do not perform well in these scenarios. One possibility to overcome this problem is to derive synchronization from an absolute time reference obtained from positioning satellites. An enhanced pilot structure may also allow resolving larger frequency misalignments between transmitter and receiver.

Other enhancements are motivated by the high vehicle densities typical in urban environments. For example, improvements to the resource allocation algorithms are necessary to improve system capacity. Sensing-based distributed resource allocation may alleviate congestion in scenarios with moderate loads, whereas centralized resource allocation may be necessary in the most challenging cases such as in traffic jams. Location information may be used to improve spatial reuse of radio resources.

For the cellular interface (e.g., used for V2N), the motivations are similar. For example, in scenarios with high densities of vehicles, multicast transmission with local distribution of the traffic may be used to alleviate network congestion. This can be realized by performing a local breakout of the traffic before it reaches the core network and redistributing the packets locally. In this way latency is minimized, which is critical for safety applications. Similarly, optimizations of scheduling protocols are necessary to reduce overhead and enable low-latency transmissions. Enhancements to the signaling protocols are also necessary to provide support for high mobility. For example, in urban scenarios, service continuity needs to be ensured for terminals that change the serving cell frequently. Other general enhancements for ITS include higher-layer protocol optimizations and security solutions for V2X.

Multimedia Broadcast and Multicast Services

Evolved multimedia broadcast multicast service (eMBMS) provides an efficient way to deliver download as well as streaming content to multiple users. Specifically, mobile video streaming will generate a major volume of network data traffic in the future. Commercial deployments of eMBMS or "LTE broadcast" are generating increasing interest, and to meet the industry and operators' demand, it is important to enhance eMBMS further, especially with a focus on use cases including linear TV, live, video on demand, smart TV, and over-the-top content.

eMBMS uses the MBMS single frequency network (MBSFN) transmission mode, where all cells in an area transmit the broadcast signal synchronously. Interference does not occur from any cell in the area where the signal arrives with a delay shorter than the cyclic prefix (CP). The CP available today for eMBMS is 16.7 μs, but this is not large enough to offer higher spectral efficiency of 2 b/s/Hz in relevant deployment scenarios such as the lower 700 and 800 MHz frequency bands and rural scenarios with smaller indoor losses or outdoor rooftop antennas for TV reception. This motivates the introduction of a longer CP, up to about 200 μs. In order to keep the relative overhead of the CP constant, the OFDM symbol length and thereby the number of subcarriers need to be increased proportionally.

Furthermore, currently only 6 subframes out of the 10 of a radio frame can be allocated to MBSFN. By extending this number, the broadcast capacity can be increased, for example, when eMBMS is deployed on a supplemental downlink (SDL) carrier. A further broadcast capacity enhancement is support for MBSFN subframes without any unicast control region. This is because with almost all subframes allocated to eMBMS, there is hardly any use for the unicast control region.

The above mentioned enhancements are being addressed in Release 14, and may ultimately lead to an LTE carrier that is dedicated to eMBMS.

Positioning Enhancements

Recently, the FCC introduced new positioning requirements with dedicated focus on indoor users [10]. It has been concluded that the baseline positioning functionalities introduced in Release 9 meet the horizontal accuracy requirements in adequately densely deployed networks [11]. The exact metric for evaluating vertical accuracy is still under discussion. Some solution components for vertical positioning were introduced in Release 13 such as terminal reporting of WiFi and Bluetooth nodes, and uncompensated barometric pressure, which will be further addressed with network assistance aspects in Release 14.

In addition, Release 14 enhancements include more general means to generate positioning reference signals (PRSs). This is motivated by the need to generate different PRSs from different remote radio heads associated with the same cell, as well as to enable better interference suppression. Moreover, the focus will also be on terrestrial beacon systems with PRS beacons on a dedicated carrier to support positioning. Furthermore, the reporting format and requirements of observed time difference of arrival measurements will be revisited to enable finer granularity reporting. Terminal receivers have become more accurate since the requirements were specified, and finer reporting prevents the positioning performance from being limited by the report quantization.

Release 14 enhancements include more general means to generate positioning reference signals (PRSs). This is motivated by the need to generate different PRSs from different remote radio heads associated to the same cell, as well as to enable better interference suppression.

Summary and Conclusion

This article has provided a high-level overview of the major technology areas considered for the evolution of LTE in Release 14, including support for reduced latency, enhancements to LTE in unlicensed spectrum, enhancements to machine-type communication, further enhancements for using multiple antennas, support for intelligent transportation systems, and enhanced support for TV services.

With the above enhancements, the LTE evolution will strive to meet the 5G requirements and address 5G use cases. As a complement to the new 5G air interface, LTE will remain an essential component of any future wireless access network.

References

[1] Ericsson Mobility Report, Nov 2015, www.ericsson.com/mobility-report.

[2] A. Osseiran *et al*., "Scenarios for 5G Mobile Wireless Communications: The Vision of the METIS Project," *IEEE Commun. Mag*., vol. 52, no. 5, May 2014, pp. 26–35.

[3] ITU-R Rec. M.2083-0, "IMT Vision Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond," Sept. 2015.

[4] 3GPP TR 38.913, "Study on Scenarios and Requirements for Next Generation Access Technologies."

[5] 3GPP TR 36.881, "Study on Latency Reduction Techniques for LTE."

[6] 3GPP TR 36.889, "Study on Licensed-Assisted Access to Unlicensed Spectrum."

[7] 3GPP TR 36.888, "Study on Provision of Low-Cost Machine-Type Communication (MTC) User equipment (UE)."

[8] 3GPP TR 36.897, "Study on Elevation Beamforming/Full-Dimension (FD) Multiple Input Multiple Output (MIMO) for LTE (Release 13)."

[9] 3GPP TR 36.885, "Study on LTE-Based V2X services."

[10] FCC 15-9 "Wireless E911 Location Accuracy Requirements, Fourth Report and Order," Feb. 2015.

[11] 3GPP TR 37.857, "Study on Indoor Positioning Enhancements for UTRA and LTE."

Biographies

Christian Hoymann received his diploma and doctoral degree in electrical engineering from RWTH Aachen University in 2002 and 2008, respectively. Since 2007 he has worked at Ericsson Research. As a principal researcher he drives the 3GPP LTE evolution, in both research and standardization. Recently, he also got involved in the development and standardization of the next generation radio access. He is currently head of Ericsson's 3GPP RAN delegation.

David Astely received his Ph.D. degree in electrical engineering from the Royal Institute of Technology, Stockholm, Sweden, in 1999. From 1999 to 2001, he was with Nokia Networks, and since 2001, he has been with Ericsson, where he has held positions in both research and development. He is currently a principal researcher at Ericsson Research working on research and standardization of future radio access.

Magnus Stattin has graduated and received his Ph.D. degree in radio communication systems from the Royal Institute of Technology in 2005. He joined Ericsson Research in Stockholm, Sweden, in June 2005. At Ericsson Research he has been working on research in the areas of radio resource management and radio protocols of various wireless technologies. He is active in concept development and 3GPP standardization of LTE, LTE-Advanced, and future wireless technologies.

Gustav Wikström received his Ph.D. in particle physics from Stockholm University in 2009. After holding a postdoctoral position at the University of Geneva, he joined Ericsson Research in 2011, where he is currently leading work to reduce user plane latency in LTE.

Jung-Fu (Thomas) Cheng is with Ericsson Silicon Valley, where he leads research and development of advanced wireless communication technologies. Since joining Ericsson in 1999, he has driven a wide range of projects evolving cellular wireless PHY and MAC layer technologies from 2.5G EDGE, 3G HSPA, 4G LTE, and 5G technologies. He holds over 100 granted U.S. patents and was named Ericsson Inventor of the Year in 2012. He was a co-recipient of the 1999 IEEE Communications Society Leonard G. Abraham Prize Paper Award in the Field of Communications Systems.

Andreas Höglund has a Master's degree in engineering physics and a Ph.D. in condensed matter theory, and started working in telecom in 2008 when he joined Ericsson Research. Since then he has been working on capacity studies, MBB traffic models for simulators, WDCMA heterogeneous networks, and power consumption reduction features and SI for MTC; has been the horizontal topic driver for massive machine-type communications (MMC) in the EU PPT research project METIS; and is currently working on standardization of eMTC and NB-IoT.

Mattias Frenne joined Ericsson in 2011 and is a senior specialist in the area of multi-antenna signal processing. He has contributed to concepts and standardization of LTE, and his current focus is on concept development and standardization of the new radio access technology in 5G. He holds an M.Sc. (1996) and Ph.D. (2002) in engineering physics and signal processing, respectively, both from Uppsala University, Sweden.

Ricardo Blasco Serrano holds M.Sc. (2007) and PhD (2013) degrees in telecommunications from the Technical University of Catalonia (UPC), Spain, and the Royal Institute of Technology, respectively. Since 2014 he has been with Ericsson Research, Kista, Sweden, working on device-to-device and vehicular communications, and as a standardization delegate in 3GPP RAN WG1.

Jörg Huschke is a master researcher in radio research at Ericsson GmbH, Germany. Since 2008, he has been working toward 3GPP standardization of MBMS for LTE and on its performance evaluation. He is active in the EBU Strategic Program Cooperating Terrestrial Networks (mobile subgroup). In 2016 he has become rapporteur for the eMBMS work item in 3GPP. He received his diploma degree in electrical engineering from Georg-Simon-Ohm University of Applied Sciences, Nuremberg, Germany in 1993.

Fredrik Gunnarsson [SM] received his M.Sc. and Ph.D. degrees in electrical engineering from Linkping University, Sweden, in 1996 and 2000, respectively. In 2001 he joined Ericsson and is now an expert in RAN Automation and Positioning at Ericsson Research as well as an associate professor at Linköping University. He is an Associate Editor of *IEEE Transactions on Vehicular Technology*. He works on concepts for positioning, as well as signal processing and control aspects of radio resource and network management.

Licensed-Assisted Access LTE: Coexistence with IEEE 802.11 and the Evolution toward 5G

Amitav Mukherjee, Jung-Fu Cheng, Sorour Falahati, Havish Koorapaty, Du Ho Kang, Reem Karaki, Laetitia Falconetti, and Daniel Larsson

The authors present a detailed overview of the design agreements for LAA, the impact of unlicensed spectrum operation on the LTE physical layer architecture, and the scope of additional enhancements beyond LTE Release 13. A range of simulations for indoor and multicarrier scenarios show that fair coexistence between LAA and WiFi can be achieved, and that deployment of LAA can provide a boost in WiFi performance.

Abstract

LAA is a new operation mode of LTE in the unlicensed spectrum, which will be featured in LTE Release 13. Under LAA, licensed carriers will be aggregated with unlicensed carriers in order to opportunistically enhance downlink user throughput while still offering seamless mobility support. In order to coexist with WiFi, some of the new functionalities required of LAA LTE include a mechanism for channel sensing based on listen-before-talk, discontinuous transmission on a carrier with limited maximum transmission duration, and multicarrier transmission across multiple unlicensed channels. This article presents a detailed overview of the design agreements for LAA, the impact of unlicensed spectrum operation on the LTE physical layer architecture, and the scope of additional enhancements beyond LTE Release 13. A range of simulations for indoor and multicarrier scenarios show that fair coexistence between LAA and WiFi can be achieved, and that deployment of LAA can provide a boost in WiFi performance.

Introduction

The proliferation of Third-Generation Partnership Project (3GPP) Long-Term Evolution (LTE) in different regions of the world demonstrates that both demand for wireless broadband data is increasing, and that LTE is an extremely successful platform to meet that demand. Existing and new spectrum licensed for use by International Mobile Telecommunications (IMT) technologies will remain fundamental for providing seamless wide-area coverage, achieving the highest spectral efficiency, and ensuring the highest reliability of cellular networks. To meet ever increasing data traffic demand (e.g., video streaming) from users and, in particular, in concentrated high traffic buildings or hotspots, more mobile broadband bandwidth will be needed. Given the large amount of spectrum available in the unlicensed bands around the globe, unlicensed spectrum is being increasingly considered by cellular operators as a complementary tool to augment their service offering. Coordinated transmission across licensed and unlicensed spectrum is also perceived to be a key feature of upcoming fifth generation (5G) radio access networks [1].

As part of this evolution, a new initiative of LTE Release 13 is the specification of licensed-assisted access (LAA) operation in the unlicensed spectrum [2, 3]. Based on the principle of carrier aggregation (CA), LAA secondary cells (SCells) carry data transmissions in the unlicensed spectrum with assistance from a primary cell (PCell) in the licensed spectrum. The PCell retains the exchange of essential control messages and also provides always available robust spectrum for real-time or delay-sensitive traffic. It enables operators to enhance the existing or planned universal seamless coverage in the LTE network with additional bandwidth and capacity. LTE Release 13 includes the specification of DL-only LAA operation as the most relevant initial use case, while uplink (UL) LAA is being incorporated into Release 14.

A key objective of the LAA feature is that the LAA design should target a fair coexistence mechanism with existing WiFi networks so as to not impact WiFi services more than another WiFi network on the same carrier would, with respect to metrics such as throughput and latency. The usage of LTE in unlicensed spectrum is a fundamental paradigm shift, since LTE physical channels have largely been designed on the basis of uninterrupted operation on licensed carriers (although Release 12 LTE added a new ON/OFF operations mode of SCells in the licensed bands).

In addition, different geographical regions have distinct regulatory requirements in terms of power spectral density for transmission in the unlicensed spectrum [2]. Therefore, Release 13 LAA targets a single global framework for LAA, with functionalities that meet regulatory requirements in different regions and bands. Furthermore, LAA design should provide sufficient configurability to enable efficient operation in different geographical regions. The LAA design should also target fair coexistence among LAA networks deployed by different operators so that the LAA networks can achieve comparable performance.

Some of the new functionalities required of LAA from a coexistence perspective include a mechanism for clear channel assessment based

Amitav Mukherjee, Jung-Fu Cheng, and Havish Koorapaty are with Ericsson Research, San Jose, California; Sorour Falahati, Du Ho Kang, and Daniel Larsson are with Ericsson Research, Stockholm, Sweden; Reem Karaki and Laetitia Falconetti are with Ericsson Research, Aachen, Germany.

0163-6804/16/$25.00 © 2016 IEEE

on listen-before-talk (LBT), discontinuous transmission (DTX) on a carrier with limited maximum transmission duration, and dynamic frequency selection (DFS) for radar avoidance in certain bands. The DTX and LBT functionalities will have a major impact on various aspects of LTE ranging from downlink physical channel design, channel state information (CSI) estimation and reporting, hybrid authomatic repeat request (HARQ) operation, to radio resource management (RRM). Prior to the LAA initiative, the coexistence of LBT-based LTE and WiFi was not evaluated in detail, with most works featuring semi-static coexistence mechanisms such as LTE almost blank subframes or time-division duplexing [4–8]. Preliminary LAA designs and coexistence evaluations corresponding to the 3GPP LAA Study Item phase were presented in [9, 10], while a Markov-chain-based analysis of LTE-WiFi coexistence with simplified LBT and no random backoff was performed in [11].

This article presents an overview of the impact of unlicensed spectrum operation and the introduction of coexistence mechanisms on the LTE physical layer framework. The LAA system architecture for downlink operation is described in detail, covering aspects such as coexistence with WiFi via channel selection and LBT, physical channel design, multicarrier operation, and RRM. Enhanced LBT algorithms are presented that further improve coexistence over the baseline LAA LBT schemes. A range of simulations for indoor and multicarrier scenarios show that fair coexistence between LAA and WiFi can be achieved, and that deployment of LAA can provide a boost in WiFi performance.

LAA Channel Access

Carrier Selection and DFS

Carrier Selection: In Japan, Europe, and the United States, between 455 and 555 MHz of unlicensed spectrum is currently available for use in the 5 GHz band. The unlicensed band can be divided into multiple carriers of 20 MHz bandwidth each. The judicious selection of one or more 20 MHz carriers with low ambient interference for operation is therefore the first step for LAA nodes to achieve good coexistence with other unlicensed spectrum deployments. However, in dense deployments with a large number of nodes, interference avoidance cannot be guaranteed through channel selection, and sharing of unlicensed carriers between different technologies is inevitable.

Carrier selection can be performed periodically in a semi-static manner since average interference levels may change in the long run due to varying numbers of neighboring nodes and traffic loads. These carriers are then configured and activated as SCells for the LAA user equipments (UEs). Carrier selection can be implemented autonomously without any specification impact by an LAA eNB by computing average received interference power estimates on candidate carriers. Additionally, UE received signal strength indicator (RSSI) measurements with configurable measurement granularity and time instances of the reports were introduced in Rel-13 LAA, and they can be a valuable tool for the assessment of hidden nodes by the evolved NodeB (eNB) near specific UEs. For example, UE measurement reports that show a high RSSI when the serving cell is inactive due to LBT can imply the presence of hidden nodes, and can be taken into account for channel (re)selection.

Dynamic Frequency Selection: DFS is a regulatory requirement for certain frequency bands in various regions, for example, to detect interference from radar systems and to avoid co-channel operation with these systems by selecting a different carrier on a relatively slow timescale. The corresponding timescales for DFS are on the order of seconds and can therefore be considered to be on an even slower timescale than carrier selection. It has been agreed in 3GPP that this functionality is an implementation issue and will not have an impact on the LTE specifications [2].

Baseline LBT Framework for a Single Carrier

The LBT procedure is defined as a mechanism by which a device performs one or more clear channel assessment (CCA) checks prior to transmitting on the channel. It is the LAA counterpart of the distributed coordination function (DCF) and enhanced distributed channel access (EDCA) medium access control (MAC) protocols in WiFi. Japanese and European regulations currently require the usage of LBT in the 5 GHz unlicensed bands, and also limit the maximum channel occupancy time for a particular transmission (e.g., 4 ms channel occupancy limit in Japan). Hence, LBT is considered to be a required functionality for fair and friendly operation in the unlicensed spectrum under a single global framework.

A straightforward approach to fair coexistence would be to make the LAA LBT procedure for both data and discovery reference signals (DRS) as similar as possible to the DCF/EDCA protocols of WiFi. This is the guiding principle behind the LAA LBT mechanism as depicted in Fig. 1, which has the following major features when energy detection (ED) is used to detect the presence of WiFi:

- Before data transmission, an LAA node must sense the medium to be idle for a *random backoff* phase comprising *N* CCA slots, where each CCA slot is of 9 μs duration. *N* is a counter drawn randomly within a dynamic contention window (CW), that is, 0 ≤ *N* ≤ CW. The *N* idle slots do not need to be contiguous in time, and the backoff counter can be decremented after each idle CCA slot.
- If the energy in a CCA slot is sensed to be above the ED threshold during random backoff, the backoff process is suspended and the counter is frozen. The backoff process is resumed, and the counter can be decremented once the medium has been idle for the duration of a defer period. A *defer period* consists of a 16 μs silent period followed by multiple CCA slots. For example, an LAA defer period of 43 μs (16 + 3 × 9 μs) is well aligned with the arbitration inter-frame space (AIFS) of EDCA best effort traffic. The backoff counter may be decremented by one after deferring is completed.
- If HARQ feedback from UEs indicates that the first subframe of the most recent DL transmission burst had 80 percent or more decoding

To support coexistence with other technologies in the unlicensed spectrum, LAA adopts dynamic carrier measurement/selection, listen-before-talk protocol, and discontinuous transmission with limited maximum duration.

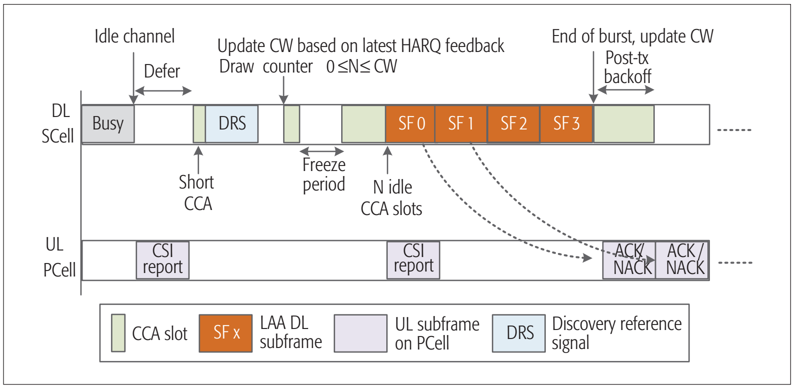


Figure 1. LAA DL transmissions with LBT CW updates based on HARQ ACK/NACK feedback. A post-transmission backoff is applied between DL bursts to prevent monopolizing the unlicensed channel. The UE provides HARQ ACK/NACK feedback and CSI reports on the licensed carrier.

errors (negative acknowledgments, NACKs), then the CW is doubled for the next LBT (up to a pre-defined maximum value such as 63 for eNBs with best-effort traffic). The CW is reset to the minimum value otherwise.

- Once the random backoff and subsequent DL transmission have been completed, a post-transmission random backoff is performed wherein a new random backoff counter is drawn and counted down before trying to transmit another DL burst, as seen in Fig. 1.
- A single short CCA period of 25 μs can be used to transmit control information without accompanying data, such as DRS.
- Four sets of minimum and maximum CW sizes, maximum channel occupancy times (MCOTs), and defer period CCA slots have been defined, corresponding to four LBT priority classes as in EDCA. For LBT class 3, CW ∈ {15, 31, 63}, and MCOT is up to 10 ms [12].

Coexistence Enhancements for LAA LBT

Coexistence with WiFi is greatly enhanced by restricting the LAA CCA starting points to LAA subframe boundaries and enforcing “freeze periods” where the backoff procedure and CCA sensing are completely suspended [13], as shown in Fig. 1. The notion of a freeze period is not a part of the LAA LBT specification, but is an implementation enhancement that can further improve coexistence. Configuring freeze periods during the LBT procedure reduces the overhead due to the possible transmission of any initial signals, since it may not be feasible to immediately start LTE data transmission at an arbitrary time instance due to the alignment of LAA SCell and PCell subframe boundaries.

As an example, the eNB may voluntarily decide to not contend for the channel for up to 11 out of the 14 orthogonal frequency-division multiplexing (OFDM) symbols (OSs) in a 1 ms period if transmission of partial subframes shorter than 11 symbols is eschewed. When coexisting with WiFi nodes that may access the channel and start transmissions at any time, the LAA eNB then essentially forfeits the channel to the WiFi nodes 78.6 percent of the time.

This is verified in Fig. 2, which shows the coexistence performance in terms of average per user throughput and outage probability of WiFi VoIP users when the LAA eNB can adapt the degree of freeze period it uses based on the observed buffer occupancy in the LAA network. The buffer occupancy metric quantifies the fraction of time there is data waiting in the eNB buffer to be served to its UEs [2]. The scenario considered here is an in-building deployment with four co-located WiFi access points (APs) and eNBs per building, along with 20 FTP users and two VoIP users per carrier per building. The performance metrics are obtained using an event-driven system simulator implemented in MATLAB with a total of 45 buildings in the simulation. For each trial (set of user drops), the simulation runtime is set to the average time needed to serve 15 files per user, and 15 such trials are conducted per traffic load point. The FTP traffic generation is based on a Poisson process model for the file arrivals, where each file is 0.5 MB in size. The maximum LAA LBT freeze period implemented here can be as large as 11 out of 14 OFDM symbols.

Figure 2 shows both per-user FTP throughput (user throughput is the average of all of its per-file throughputs) and VoIP outage metrics as a function of the total served traffic, which increases as the file arrival rate per user is increased. Here, VoIP users with 98th percentile latency greater than 50 ms are considered to be in outage (out of a total of 90 VoIP users in the simulation). The figures clearly show that good coexistence with WiFi is possible in the indoor scenario, even when a conservative ED threshold of –62 dBm is used by LAA. For example, the mean per user throughput in the WiFi network of operator A is increased by around 40 percent in both DL and UL when coexisting with LAA for the same served traffic level of 8 Mb/s.

Multicarrier LBT

Simultaneous operation on multiple unlicensed channels or carriers is a key technique for maximizing the data delivered during a transmission opportunity. As an example, IEEE 802.11ac supports transmission bandwidths of up to 160 MHz, which would span eight contiguous 20 MHz unlicensed channels in the 5 GHz band. The design of a multicarrier operation mode for LAA with concurrent transmission on multiple unlicensed SCells should continue to adhere to the principle of fair coexistence with WiFi, while being able to quickly detect transmission opportunities across multiple channels.

With regard to coexistence, a brief overview of the multicarrier LBT procedure in WiFi is provided next. WiFi adopts a hierarchical channel bonding scheme by combining contiguous 20 MHz sub-channels in a non-overlapping manner. One of these contiguous sub-channels is designated as a primary channel on which a complete random backoff cycle is performed, while the others are designated as secondary channels. Counting down of the random backoff counter is based only on the outcome of clear channel assessments on the primary channel. On the secondary channels, only a quick CCA check is performed for point coordination function interframe space (PIFS) duration (generally 25 μs) before the potential start of transmission,

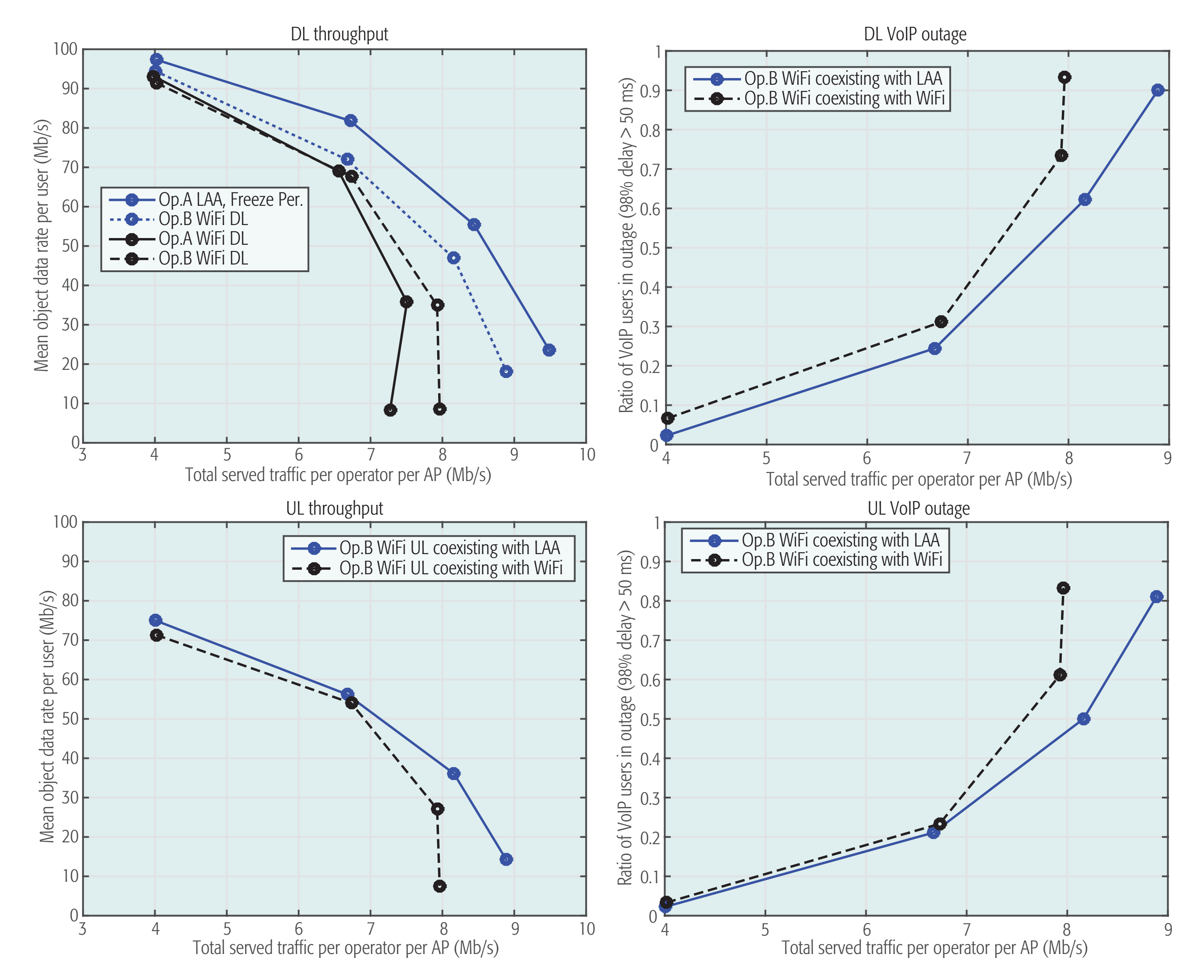


Figure 2. 3GPP indoor single-carrier deployment showing the improvement in Wi-Fi FTP and VoIP performance when LAA LBT employs freeze periods together with an ED threshold of –62 dBm per CCA slot. Blue lines indicate Wi-Fi-LAA coexistence, while black lines indicate WiFi-WiFi coexistence scenarios. LAA does not utilize the licensed carrier for data. Operator A's network has only DL traffic, and operator B network has both DL and UL FTP traffic with 80/20 split, along with 20 FTP users and two VoIP users per carrier per building. For both WiFi and LAA, transmit bursts are of 4 ms duration, 256-quadrature modulation (QAM) is supported, and antenna configuration is 2 Tx–2 Rx [2].

to determine which of the secondary channels are also available in addition to the primary. The final transmission therefore always includes the primary channel. Upon expiration of the backoff counter, the overall transmission bandwidth (20 MHz, 40 MHz, 80 MHz, or 160 MHz) is determined by the results of the secondary CCA checks. The signal and energy detection thresholds for secondary channels are generally higher than those for the primary channel, and scale up with increasing channel bandwidth.

Thus, two main alternatives for LAA multicarrier LBT are apparent.

- **Alt. 1: Single Random Backoff Channel:** Similar to WiFi, only one full-fledged random backoff (as defined above) needs to be completed on any one carrier, along with quick CCA checks on the other channels, before transmission occurs.
- **Alt. 2: Parallel Random Backoff Channels:** Multiple SCells need to each have individually completed full-fledged random backoffs before transmitting simultaneously.

Both alt. 1 and alt. 2 are supported in Release 13 LAA. Representative examples of these multicarrier LBT alternatives are compared in Fig. 3 for a scenario with three LAA SCells that are assigned a common random backoff counter. In the case of alt. 1, SCell 1 finishes counting down first and is designated as the channel with the full-fledged random backoff procedure. To determine whether any other channels are eligible for transmission, the most recent slots of the random backoff procedure corresponding to these channels are examined, and the channels that are found to have been idle for the duration of a PIFS are also used for transmission, which is SCell 3 in Fig. 3. In the case of alt. 2, all SCells that finish their countdown before a predefined wait limit (defined in terms of CCA slots) transmit simultaneously.

A performance evaluation for multi-car-

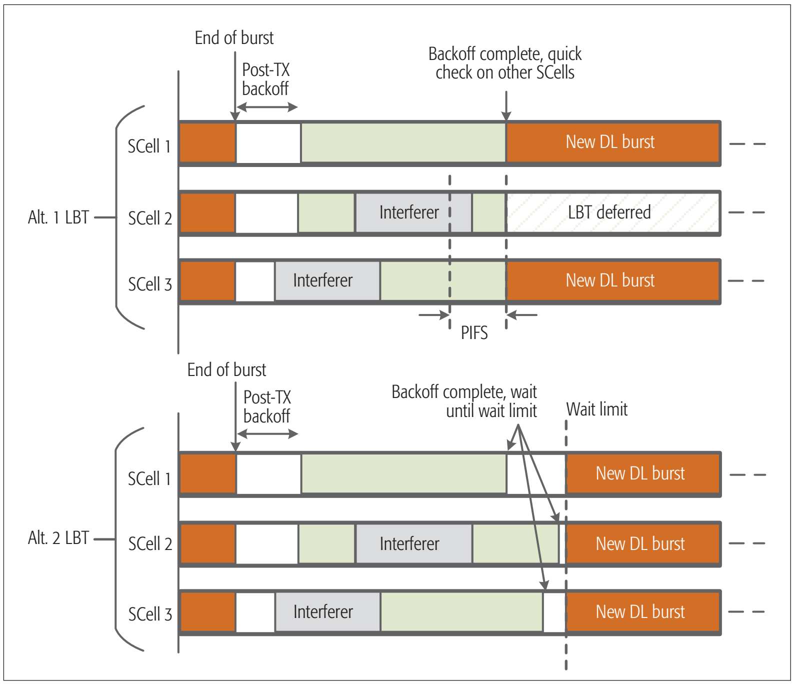


Figure 3. Comparison of LAA multicarrier LBT access schemes: alt. 1, with a single random backoff channel, and alt. 2, with multiple random backoff channels. In the alt. 2 example, a static prespecified wait limit is defined; SCells that have completed backoff before the wait limit defer transmission accordingly.

rier LBT over 80 MHz is shown in Fig. 4. The overall system performance results clearly show that from the coexistence point of view and the impact on the non-replaced WiFi network, both classes of multi-channel LAA LBT schemes are viable and can increase the performance of a multi-carrier WiFi network compared to when it is coexisting with another WiFi network. Also, alt. 1 with a single random backoff channel offers better coexistence due to the more agile random backoff channel selection and better alignment with the WiFi procedure.

Downlink LAA Framework

Physical Channels and Partial Subframes

A summary of the major differences in physical channel and RS design between Release 13 LAA and Release 12 LTE is shown in Table 1. The design of the data-bearing physical downlink shared channel (PDSCH) was one of the major focus areas of Release 13 LAA. One of the main issues regarded the usage of partial subframes that occupy fewer than 14 OFDM symbols (a 1 ms TTI) at the beginning or end of DL bursts. At the start of a burst, a partial subframe can be useful since the starting point for data transmission varies due to the random completion time of LBT. Furthermore, the end of a DL burst may have to be truncated into a partial subframe due to regulatory restrictions on maximum channel occupancy.

In practice, it is difficult to implement a large number of different starting symbol positions at the beginning of a DL burst. This is because of the multiple steps involved in preparation of a data subframe, ranging from scheduling, layer 2 control processing, encoding, scrambling and modulation, and transport of digital samples to remote radio heads. Furthermore, preemptively preparing different subframe versions for different possible starting points raises costs due to increased memory requirements. Ultimately, it was agreed to support up to two starting points (either the first or eighth OFDM symbol) at the start of a DL burst, and multiple partial subframe lengths (from 3 to 12 OFDM symbols) for the last subframe of a burst. The LBT scheme with freeze periods described earlier has a natural synergy with limiting the starting positions at the initiation of a burst: if LBT does not succeed until the first/eighth symbol, the next CCA is deferred until the end of that subframe/slot. If the LAA eNB is unable to start PDSCH transmission immediately after clearing LBT, a short initial signal, for example, comprising primary and secondary synchronization sequences, may optionally be transmitted before PDSCH transmission begins.

The length of a partial subframe at the end of a DL burst will be indicated both in the partial and previous subframes using cell-specific physical downlink control channel (PDCCH) signaling. PDSCH in LAA will support most of the single-codeword and dual-codeword transmission schemes in LTE Release 12, with the exception of multi-user multiple-input multiple-output (MIMO), closed-loop unit-rank spatial multiplexing, and single antenna port beamforming (transmission modes 5, 6, and 7). Since the motivation is to use LAA as a throughput booster, it is reasonable to focus on the MIMO transmission modes with the highest spectral efficiency, such as modes 9 and 10. To facilitate detection of bursts and fine synchronization, every DL subframe will carry at least one symbol of cell-specific reference signals (CRSs).

With a single starting position at subframe boundaries together with LBT with freeze periods described above, there is no impact on UE-specific demodulation reference signals (DMRSs) and channel state information reference signals (CSI-RSs) in the first subframe of a burst since they are located after the fifth OS within a subframe. For the same reasons, no changes are required for control channel design when DL bursts start from the first OS, with regard to PDCCH and the enhanced PDCCH (EPDCCH). If a 7-symbol partial subframe is used at the start of a burst, the PDCCH and PDSCH resource mapping is the same as in the first slot of a regular full-length subframe, while the start symbol of the EPDCCH is offset by 7 OFDM symbols. For the partial subframe of various lengths at the end of a burst, the reference signal mappings are based on existing mappings defined for the downlink pilot time slot (DwPTS) used in TDD special subframes.

Finally, the physical multicast channel (PMCH) and physical broadcast channel are not included in Release 13 LAA, since single-frequency operation and transmission without a licensed carrier are not in the scope of the feature.

Control Signaling

In Release 10 CA, an SCell may carry scheduling grants for UEs served on that same SCell (referred to as self-scheduling), or UEs on a par-

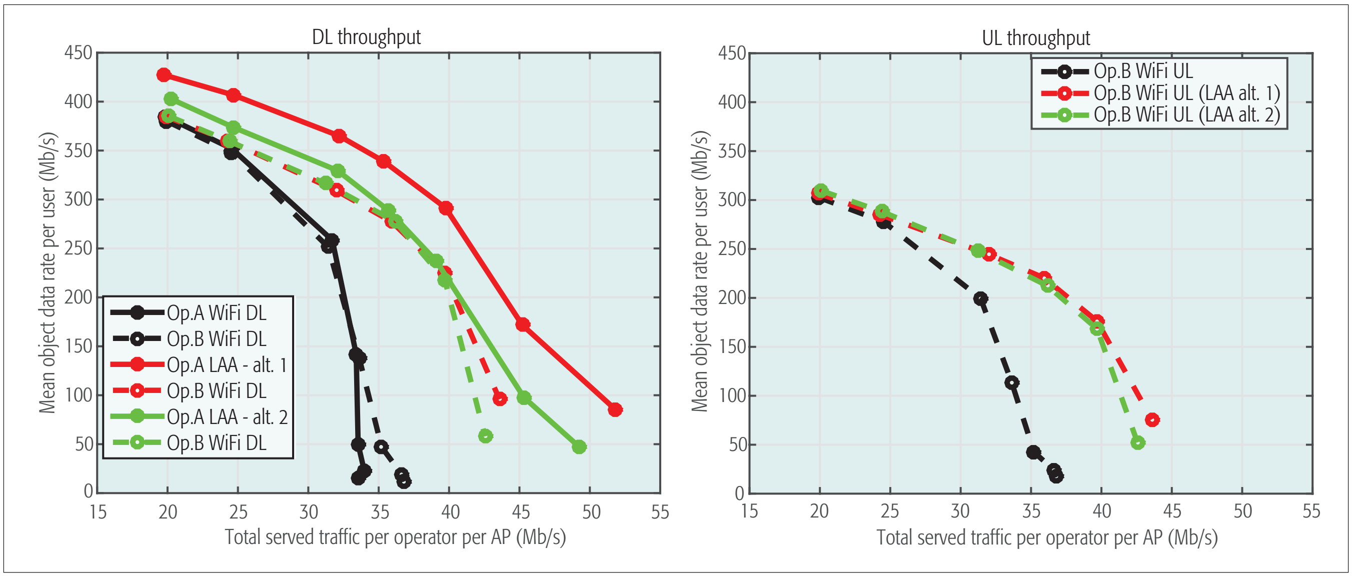


Figure 4. Mean user throughput vs. served traffic per AP per operator for the indoor multi-carrier deployment scenario with FTP traffic using up to 80 MHz transmission bandwidth. The non-replaced WiFi network is operator B. Left and right plots correspond to DL and UL per user throughput results, respectively.

ticular SCell may be scheduled from the PCell or another SCell via cross-carrier scheduling configuration. To support self-scheduling on the LAA SCell, either the PDCCH or the EPDCCH can be used to send DL resource assignments. Therefore, the physical control format indicator channel (PCFICH) may also be transmitted on LAA SCells to indicate how many OFDM symbols are allocated for PDCCH in a subframe. However, as discussed earlier, the physical HARQ indicator channel (PHICH) will not be used to convey HARQ ACK/NACKs for UL transmissions. Both PCFICH and PHICH resources will continue to span the system bandwidth in the first OFDM symbol of each DL subframe, as in Release 12 LTE.

Following the LTE CA framework, the physical uplink control channel (PUCCH) carrying HARQ-ACK and CSI for all aggregated cells should be sent on the UL PCell. Since the PUCCH resources on the PCell are reserved and always available (unlike the LAA SCell), sending UCI on the PUCCH is one important advantage of the LAA LTE PHY layer over the WiFi PHY layer. The LTE PUCCH is designed for coverage and reliability via very low rate coding (rate-1/3 convolutional coding), repetition, and frequency hopping. Transmitting on the licensed PCell further allows the UE to transmit at higher powers and at lower carrier frequencies (with lower path losses). WiFi control frames for both UL and DL utilize binary phase shift keying (BPSK) rate-1/2 convolutional coding, and the frame acknowledgment design requires substantially more bits than the LTE design, although the gap between data transmission and ACK reception is significantly lower for WiFi. In all, the WiFi control frame has been designed to provide coverage in localized areas such as indoor deployment and is less suitable for outdoor deployment. The LTE control channel design in comparison enables reliable outdoor deployment and a larger coverage area.

Component	Release 13 LAA	Release 12 LTE
PDSCH	Can start at slot boundaries	Start at subframe boundaries
PDCCH	Up to first three symbols of a slot	Up to first four symbols of a subframe
EPDCCH	Can occupy one slot	Can occupy last 11 OS
PHICH	Not used for UL HARQ feedback	Used for UL HARQ feedback
PBCH/PMCH	Not supported	Supported
CRS	1- or 2-symbol CRS in up to 8 out of 10 subframes per frame allowed	1- or 2-symbol CRS in up to 6 out of 10 subframes per frame allowed
PSS/SSS	Can appear outside subframe 0/5	Appear only in subframe 0/5

Table 1. Summary of PHY changes in Release 13 LAA.

Radio Resource Management and CSI Measurements

The combination of the LBT and maximum transmission burst duration functionalities of LAA implies that LTE reference signals are not guaranteed to be transmitted with a fixed periodicity on LAA SCells. This can affect the methods by which RRM, CSI measurements and feedback, and time-frequency tracking are currently supported in LTE. Similarly, RRM measurements form the basis for cell selection and mobility management, and closed-loop link adaptation is not feasible without accurate CSI measurements and feedback.

In Release 12, periodic transmission of primary/secondary synchronization sequences (PSS/SSSs), CRSs, and CSI-RSs are generally used to achieve these objectives. Since periodic reference signal transmission is no longer feasible on LAA SCells, this raises the question if PCell reference signals can be utilized for at least coarse time-frequency synchronization and automatic gain control (AGC) adjustment on the SCells. While coarse timing synchronization may be possible using the PCell RS in a co-located scenario, AGC

To allow additional deployment scenarios for LTE that utilize unlicensed spectrum, it is important to extend the applicable deployment scenarios beyond those enabled by carrier aggregation with licensed spectrum and the associated stringent time synchronization requirements.

adjustment would not be feasible due to the PCell potentially operating on a carrier (e.g., 2 GHz) that has substantially different characteristics and path loss compared to LAA SCells in the 5 GHz band. This difference in long-term channel properties also rules out using PCell RS for channel estimation filter adjustment on LAA SCells.

Based on the above discussion, the Release 13 solution is to transmit constant-power discovery reference signals (DRSs) on the LAA SCells, subject to a single CCA duration of 25 μs without random backoff. The signals comprising the LAA DRSs are the same as symbols 0–11 of the Release 12 DRS, which comprised PSS/SSS/CRS and was designed for small cell PHY enhancements. The transmission burst containing LAA DRSs cannot exceed 1 ms in duration, and will be attempted to be sent periodically every 40 ms or so. UEs will be configured with a discovery measurement timing configuration (DMTC) of six subframes, within which they can attempt to detect and measure DRSs of serving and adjacent LAA cells.

Moreover; CSI-RSs together with CSI-IMs (or other known unused resource elements) can be used to derive CSI reports from the UE when they are available. Due to the unpredictable availability of the unlicensed carrier, the most practical approach would be to rely only on aperiodic CSI reports for the LAA SCell, as opposed to periodic CSI reports. Furthermore, the CRS and CSI-RS power will be the same within a burst but may be varied across transmission bursts; therefore, UEs should not average CRS/CSI-RS measurements across bursts.

Evolution beyond Release 13

The initial DL-only LAA framework developed in Release 13 is amenable to several potential enhancements in future releases in order to create a full-fledged LAA design with both DL and UL transmissions and support for aggregation of a large number of unlicensed channels.

Uplink LAA

Several aspects of UL LAA, such as HARQ and UL LBT, were discussed during Release 13. It was agreed that UL HARQ should follow an asynchronous protocol, similar to LAA DL HARQ. In other words, UL retransmissions are explicitly rescheduled by the eNB, as opposed to automatic retransmission 4 ms after an eNB NACK as in Release 12.

With regard to the framework of UL LBT, it was recognized that UL LBT imposes an additional LBT step for UL transmissions that were scheduled by an LAA SCell (self-scheduling), since the UL grant itself requires a DL LBT by the eNB. Therefore, Release 13 LAA recommended that the UL LBT for self-scheduling should use either a single CCA duration of at least 25 μs (similar to DL DRS) or a random backoff scheme with a defer period of 25 μs including a defer duration of 16 μs followed by one CCA slot, and a maximum contention window size between three and seven.

The exact specification of UL LBT and enhancements to UL scheduling, PUCCH design, and data transmission are expected to be finalized in a Release 14 work item.

LAA with 32 Carriers

Wideband transmissions are a key feature for enabling high user data rates, and this is especially true as we evolve toward 5G. As discussed earlier, IEEE 802.11ac currently supports transmission bandwidths of up to 160 MHz, and further improvements may be made in 802.11ax. In contrast, Release 13 LAA can aggregate up to 100 MHz on the downlink by aggregating five DL carriers. Therefore, LAA should be enhanced to support system bandwidths similar to 802.11ac in unlicensed spectrum. In Release 13, a separate 3GPP work item specified aggregation of up to 16 or 32 carriers, which is a natural candidate for application to LAA. With 32 aggregated carriers, LAA will then be able to support a transmission bandwidth of 640 MHz to a single UE. This would impact a number of PHY-layer aspects, such as the need to support PUCCH on LAA SCells to reduce control overhead on the PCell, and enhancements in scheduling with such a large number of available carriers.

Dual Connectivity Support

To allow additional deployment scenarios for LTE that utilize unlicensed spectrum, it is important to extend the applicable deployment scenarios beyond those enabled by carrier aggregation with licensed spectrum and the associated stringent time synchronization requirements. It is therefore proposed to extend the design to also allow for dual connectivity (DC) operation between LTE in licensed and unlicensed spectrum with non-ideal backhaul in Release 14, which would have much looser synchronization requirements compared to CA. Supporting DC would necessitate the introduction of PUCCH and random access channels on the LAA UL, in addition to changes in radio link monitoring of the unlicensed secondary carriers.

Conclusions

This article presents an overview of licensed assisted access in Release 13 LTE for operation in unlicensed spectrum. It is shown how the introduction of new functionalities such as DTX and LBT necessitates numerous changes in the DL physical channels, HARQ feedback procedures, scheduling, RRM mechanisms, and CSI acquisition. Detailed system-level simulation results are presented to show that fair coexistence between LAA and WiFi can be achieved in a range of single-carrier and multi-carrier scenarios. Finally, an overview of desirable enhancements for LAA in future LTE releases was presented.

References

[1] E. Dahlman *et al.*, "5G Wireless Access: Requirements and Realization," *IEEE Commun. Mag.*, vol. 52, no. 12, Dec. 2014, pp. 42–47.

[2] 3GPP TR 36.889 v13.0.0, "Feasibility Sudy on Licensed-Assisted Access to Unlicensed Spectrum," July 2015; www.3gpp.org

[3] RP-151045, "New Work Item on Licensed-Assisted Access to Unlicensed Spectrum," 2015; www.3gpp.org

[4] F. Liu *et al.*, "Small Cell Traffic Balancing over Licensed and Unlicensed Band," *IEEE Trans. Vehic. Tech.*, vol. 64, no. 12, Dec. 2015, pp. 5850–65.

[5] E. P. L. Almeida *et al.*, "Enabling LTE/WiFi Coexistence by LTE Blank Subframe Allocation," *Proc. IEEE ICC*, 2013.

[6] A. M. Cavalcante *et al.*, "Performance Evaluation of LTE and WiFi Coexistence in Unlicensed Bands," *Proc. IEEE VTC-Spring*, 2013.

[7] N. Rupasinghe and I. Guvenc, "Licensed-Assisted Access for WiFi-LTE Coexistence in the Unlicensed Spectrum," *Proc. IEEE GLOBECOM Wksp. Emerging Tech. for 5G*, Austin, TX, Dec. 2014.

[8] H. Zhang *et al.*, "Coexistence of WiFi and Heterogeneous Small Cell Networks Sharing Unlicensed Spectrum," *IEEE Commun. Mag.*, vol. 53, no. 3, Mar. 2015, pp. 158–64.

[9] A. Mukherjee *et al*., "System Architecture and Coexistence Evaluation of Licensed-Assisted Access LTE with IEEE 802.11," *Proc. IEEE ICC Wksp. LTE-U*, London, U.K., June 2015.

[10] J. Jeon *et al*., "LTE with Listen-Before-Talk in Unlicensed Spectrum," *Proc. IEEE ICC Wksp. LTE-U*, London, U.K., June 2015.

[11] C. Chen, R. Ratasuk, and A. Ghosh, "Downlink Performance Analysis of LTE and WiFi Coexistence in Unlicensed Bands with aSimple Listen-Before-Talk Scheme," *Proc. IEEE VTC-Spring*, Glasgow, Scotland, May 2015.

[12] 3GPP TS 36.300, 2016; www.3gpp.org

R1-151131, "Further Details on LBT Design for DL," *Ericsson*, Mar. 2015; www.3gpp.org.

Biographies

Amitav Mukherjee received his B.S. degree from the University of Kansas, Lawrence, in 2005, his M.S. degree from Wichita State University, Kansas, in 2007, both in electrical engineering, and his Ph.D. degree in electrical and computer engineering from the University of California, Irvine, in 2012. He is currently a research engineer at Ericsson Research, San Jose, California, where he is involved with the standardization of LTE in unlicensed spectrum. From 2012 to 2014, he was a 3GPP RAN1 delegate at Hitachi America Ltd., Santa Clara, California. His research interests encompass statistical signal processing and wireless communications, with over 65 refereed publications and 50 pending/issued patents in these areas.

Jung-Fu (Thomas) Cheng is with Ericsson Silicon Valley where he leads research and development of advanced wireless communication technologies. Since joining Ericsson in 1999, he has driven a wide range of projects evolving cellular wireless PHY and MAC layer technologies from 2.5G EDGE, 3G HSPA, 4G LTE, to 5G technologies. His research interests include iterative processing, coding, and signal processing algorithms for wireless communications. He was the principal contributor to the LTE turbo and convolutional coding and rate matching procedures. He is responsible for driving Ericsson's global strategy, research, and standardization of small and heterogeneous cell operations using LTE technologies. He holds over 100 granted U.S. patents and was named Ericsson Inventor of the Year in 2012. He was a co-recipient of the 1999 IEEE Communications Society Leonard G. Abraham Prize Paper Award in the Field of Communications Systems. He received his B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei. He received his Ph.D. degree in electrical Engineering from California Institute of Technology, Pasadena.

Sorour Falahati received her B.S. degree in electrical engineering from Sharif University of Technology, Tehran, Iran, in 1994, and both her M.Sc. and Ph.D degrees from Chalmers University of Technology, Gteborg, Sweden, in 1996 and 2002, respectively. From January 2003 until October 2005, she worked in the Signals and Systems Group, Uppsala University, Sweden. Since November 2005 she has been with Ericsson Research in Stockholm, Sweden, and has worked on a range of topics including signal processing for communication systems, control channel design, wireless backhaul systems, and spectrum sharing in unlicensed bands. Currently, she is involved in standardization efforts in 3GPP for evolution of LTE, including operation of LTE in unlicensed spectrum, and in ETSI BRAN for development of a harmonized standard for operation in unlicensed spectrum at 5 GHz.

Havish Koorapaty (Havish.Koorapaty@ericsson.com) received his B.S., M.S., and Ph.D. degrees in electrical and computer engineering from North Carolina State University in 1991, 1993, and 1996, respectively. He has been with Ericsson Research since 1996, where he has worked in the general area of wireless communications systems including cellular and satellite systems. His work spans a wide range of topics including error control coding, location determination and tracking, mobile phone systems engineering, broadband wireless system design, wireless backhaul solutions, energy efficiency, spectrum sharing, and small cells. He has over 100 technical papers and patents in these areas. Recently, he has worked on LTE evolution and has been involved in standardization efforts in 3GPP for LTE including serving as Rapporteur for the licensed-assisted access study and work items. He received the 3GPP Excellence award for his contributions to 3GPP in 2015.

Du Ho Kang (du.ho.kang@ericsson.com) received his M.S. degree from Seoul National University, Korea, in 2010 and obtained his Ph.D. degree in 2014 in the area of radio communication systems from KTH Royal Institute of Technology, Sweden. He joined Ericsson Research as an experienced researcher and has been engaged in the standardization and regulation of LTE evolution and 5G. He is also serving as the Ericsson delegate of METIS-II, the EU 5G flagship project. His general interests include spectrum technologies and regulations for future wireless access networks.

Reem Karaki is a research engineer in Radio Protocols at Ericsson Germany. Her research interests include LTE licensed assisted access (LAA), LTE/WiFi aggregation, and dual connectivity. She received an M.Sc. in communications engineering from RWTH, Germany, in 2014.

Laetitia Falconetti holds an M.Sc. from Karlsruhe University and a Ph.D. from RWTH). She joined Ericsson Research in 2008 and is a senior researcher in the Radio Network Algorithms Group at Ericsson Research, Kista, Sweden. Her research interests include radio resource management, energy-efficient mobile communications, and, more generally, LTE evolution with features such as LAA and latency reduction.

Daniel Larsson received his M.Sc.E.E. from the Royal Institute of Technology in 2006. He joined Ericsson in 2007, and has been working on the development and standardization of mobile broadband technologies since 2008. His research has included work in the areas of carrier aggregation, mobility, positioning, broadcast techniques (eMBMS), TDD technologies, and interference mitigation techniques. He is currently working on operation of LTE in unlicensed spectrum and latency reduction in LTE. He is also head of the delegation at Ericsson in 3GPP RAN WG1 for 4G and 5G technologies. He received the Ericsson Inventor of the Year award in 2015.

BIO-INSPIRED CYBER SECURITY FOR COMMUNICATIONS AND NETWORKING

Wojciech Mazurczyk

Sean Moore

Errin W. Fulp

Hiroshi Wada

Kenji Leibnitz

Nature is Earth's most amazing invention machine for solving problems and adapting to significant environmental changes. Its ability to address complex, large-scale problems with robust, adaptable, and efficient solutions results from many years of selection, genetic drift, and mutations. Thus, it is not surprising that inventors and researchers often look to natural systems for inspiration and methods to solve problems in human-created artificial environments. This has resulted in the development of evolutionary algorithms including genetic algorithms and swarm algorithms, and of classifier and pattern detection algorithms, such as neural networks, for addressing hard computational problems.

A natural evolutionary driver is to survive long enough to create a next generation of descendants and ensure the survival of the species. One factor in survival is an organism's ability to defend against attackers, both predators and parasites, and against rapid changes in environmental conditions. Analogously, networks and communications systems use cyber security to ensure the survival of their assets against cyber criminals, hostile organizations, hackers, activists, and sudden changes in the network environment. Many of the defense methods used by natural organisms may be mapped to cyber space to implement effective cyber security. Some examples include immune systems, invader detection, friend vs. foe, camouflage, and mimicry. Many cyber security technologies and systems in common use today have their roots in bio-inspired methods, including anti-virus, intrusion detection, threat behavior analysis, attribution, honeypots, counterattack, and the like. As the threats evolve to evade current cyber security technologies, similarly bio-inspired security and defense technologies evolve to counter the threat.

An objective of this Feature Topic was to survey current work in the area of bio-inspired cyber security for communications and networking. Hopefully this Feature Topic provides the ComSoc community a better understanding of the current evolutionary state of cyber threats, defenses, and intelligence, and helps the community plan for future transitions of this research into practical implementations.

In this Feature Topic, we are delighted to present a selection of four articles that contribute to the enhancement of knowledge in bio-inspired cyber security for communications and networking. The collection of these high-quality articles provides a view on the latest research advances in this field.

The first article, "Hive Oversight for Network Intrusion Early-Warning Using DIAMoND (HONIED): A Bee-Inspired Method for Fully Distributed Cyber Defense" by Korczynski *et al.*, describes an interesting concept of a self-organized anomaly detection system. This bio-inspired approach is derived from the social interactions of honey bees observed in nature. In honey bee foraging, system participants do not define the search target a priori; instead, participants identify anomalies (resources) as they encounter them. The foraging methods may be mapped to computer system networks in order to detect and mitigate distributed attacks in an automated fashion.

In the second article, "Bio-Inspired Cybersecurity for Wireless Sensor Networks," Bitam *et al.* first carefully review existing bio-inspired techniques developed for improving cyber security of cyber-physical systems using wireless sensor networks. The authors propose a generic bio-inspired machine-learning model called Swarm Intelligence for WSN Cybersecurity (SIWC) that addresses drawbacks of prior bio-inspired solutions. SIWC is a neural network system trained by swarm intelligence optimization to automatically and efficiently determine the optimal critical parameters used to detect cyber-attacks.

In the next article "Decapitation via Digital Epidemics: A Bio-Inspired Transmissive Attack," Chen *et al.* address an emerging attack pattern called transmissive attack in which an attacker leverages diverse communication paths to approach the target and performs malicious activities. The authors provide an overview of commonly used epidemic models for communication systems and then relate transmissive attacks to these epidemic models. Simulations and experiments in mobile social networks demonstrate the utility of epidemic models for assessing the trade-off between the transmissive attacks' success rate vs. the risk of exposure (detection).

Finally, in "Bio-Inspired RF Steganography via Linear

Chirp Radar Signals," Zhang *et al.* propose RF steganography for concealing communication information in a linear chirp radar signal. The main idea of the described approach is inspired by the chirping sound made by birds. Some bird calls and bird songs can convey messages that are only partially understood by other species, and some information will only be picked up by birds of the same species or even between only a few individuals.

We hope readers enjoy this Feature Topic and find the articles interesting. In addition, we also hope that the presented results stimulate further research in these important areas of information and network security.

We would also like to express our thanks for the support and help of Osman S. Gebizlioglu, Editor-in-Chief of *IEEE Communications Magazine*, Joseph Milizzo from the Communications Society staff, the leading researchers contributing to the Feature Topic, and the excellent reviewers for their great help and support that made this Feature Topic possible.

Biographies

Wojciech Mazurczyk [M'11, SM'13] (wmazurczyk@cygnus.tele.pw.edu.pl) received his M.Sc., Ph.D. (Hons.), and D.Sc. (habilitation) degrees in telecommunications from the Warsaw University of Technology (WUT), Poland, in 2004, 2009, and 2014, respectively. He is currently an associate professor with the Institute of Telecommunications at WUT. His research interests include bioinspired cybersecurity and networking, information hiding, and network security. Since 2013, he has been an Associate Technical Editor of *IEEE Communications Magazine*.

Sean Moore [M'01, SM'03] is the chief technology officer and vice president of research for Centripetal Networks. In the past, he was a chief scientist and chief architect at Avaya, chief scientist at Cetacean Networks, and senior director of R&D at MadeToOrder.com. He is a former Editor-in-Chief of *IEEE Communications Magazine*. He received a B.S. degree in electrical engineering from Tulane University in 1983, an M.S. in mathematics from the University of New Orleans in 1990, and M.S. and Ph.D. degrees in computer science from Dartmouth College in 1993 and 1994, respectively.

Errin W. Fulp holds an M.Sc. (1994) in computer science and a Ph.D. (1999) in computer engineering from North Carolina State Universit, and is a full professor at Wake Forest University. He is an author of over 55 publications, 3 patents, and 20 invited talks on computer security and networks. He was involved in one startup company. His main research interests are moving target system defenses, agent-based (swarm intelligence) network security and management, and bio-inspired security methods for the smart grid.

Hiroshi Wada holds a Ph.D. (2009) in computer science from the University of Massachusetts Boston. He is a staff engineer with Unitrends Australia, and leads research and development. Before joining Unitrends he was a senior researcher with NICTA and a VP research at Yuruware, which was a startup incubated in NICTA subsequently acquired by Unitrends in 2014. His research interests include distributed computing, dependable system operation, performance engineering, and search-based software engineering. In these areas he has published over 30 refereed papers.

Kenji Leibnitz received M.Sc. and Ph.D. degrees in information science from the University of Würzburg, Germany. Since 2010 he has been a senior researcher at the National Institute of Information and Communications Technology and a guest associate professor at Osaka University; since 2013 he has been with the Center of Information and Neural Networks. His research interests include modeling and performance analysis of communication networks, especially the application of biologically and brain inspired mechanisms to self-organization in future networks.

Hive Oversight for Network Intrusion Early Warning Using DIAMoND: A Bee-Inspired Method for Fully Distributed Cyber Defense

Maciej Korczyński, Ali Hamieh, Jun Ho Huh, Henrik Holm, S. Raj Rajagopalan, and Nina H. Fefferman

The authors investigate the potential for a self-organizing anomaly detection system inspired by those observed naturally in colonies of honey bees. They provide a summary of findings from a recently presented algorithm for a nonparametric, fully distributed coordination framework that translates the biological success of these methods into analogous operations for use in cyber defense and discuss the features that inspired this translation.

Abstract

Social insect colonies have survived over evolutionary time in part due to the success of their collaborative methods: using local information and distributed decision making algorithms to detect and exploit critical resources in their environment. These methods have the unusual and useful ability to detect anomalies rapidly, with very little memory, and using only very local information. Our research investigates the potential for a self-organizing anomaly detection system inspired by those observed naturally in colonies of honey bees. We provide a summary of findings from a recently presented algorithm for a nonparametric, fully distributed coordination framework that translates the biological success of these methods into analogous operations for use in cyber defense and discuss the features that inspired this translation. We explore the impacts on detection performance of the defined range of distributed communication for each node and of involving only a small percentage of total nodes in the network in the distributed detection communication. We evaluate our algorithm using a software-based testing implementation, and demonstrate up to 20 percent improvement in detection capability over parallel isolated anomaly detectors.

Introduction

Over the past years, cyber-attackers have taken advantage of the massive acceleration in the adoption of virtualization and cloud computing, the Internet of Things (IoT), and mobile devices as an increase in potential targets and expanding attack surface. Motivations are the major characteristics that differentiate malicious actors. Organized crime is interested in economic gain, nation-states are mostly interested in cyber-espionage, whereas hacktivists can be motivated politically or ideologically.[1] Cyber-attack strategies have also evolved significantly: modern malicious activities are spread stealthily over a large number of malicious machines. Those can be compromised or rented from so-called bulletproof hosting providers that ignore all abuse notifications [1]. This increases the chance of cyber-criminal success, either decreasing the probability the attack will be noticed or launching a distributed denial of service (DDoS) attack as a smokescreen to cover virus or malware installation, and/or financial or data theft.[2]

To address these more challenging types of cyber-attacks, recent defenses have introduced the idea of sharing information across organizational boundaries, allowing collaboration to achieve rapid detection and mitigation for a variety of cyber-attacks, especially those for which prior knowledge is scant or nonexistent. Indeed, an entire new infrastructure is being created with new sharing protocols, cyber threat “exchanges,” and government backing. Automated cyber data processing and sharing is already being promoted as the new defensive strategy against smart and highly distributed adversaries. However, there are some fundamental challenges to address before this paradigm can become reality, such as:

- Policy issues that prevent sensitive data from being shared between organizations
- System scalability
- Semantics of the data being exchanged
- Alert correlation

As cyber-attacks are evolving rapidly, the data captured in one particular environment may be incomparable to data from another, vitiating any gains from sharing. Any form of detection that relies on comparison of semantically rich data is thus in jeopardy if the data comes from sensors in different domains. Even if direct comparison is possible, it is not guaranteed that the existing alert correlation techniques will be able to reconstruct novel, complex attack scenarios.

Honey Bees as an Evolved Anomaly Detection Machine

Colonies of honey bees rely on foraging workers to discover and share locations of flowering plants from which to gather the pollen and nectar used for food. The colony operates under many time-varying constraints: different plants flower at different times of year and/or day, other ani-

[1] http://www.mcafee.com/mx/resources/reports/rp-quarterly-threats-aug-2015.pdf

[2] https://www.neustar.biz/resources/whitepapers/ddos-protection/2014-annual-ddos-attacks-and-impact-report.pdf

Maciej Korczyński is with Delft University of Technology; Ali Hamieh and Nina H. Fefferman are with Rutgers University; Jun Ho Huh and S. Raj Rajagopalan are with Honeywell ACS Labs; Henrik Holm is with Forest Glen Research, LLC.

0163-6804/16/$25.00 © 2016 IEEE

mals also eat the plants/flowers, or the nectar and pollen are depleted by both direct competition with other insects/bees and by their own colony mates having already gathered the resources, making additional trips redundant. Each of these challenges must be met efficiently since the rate of resource acquisition determines the probability of colony growth, reproduction, and survival through the winter [2]. Meeting these challenges requires the colony (using only the relatively simple cognition and communication available to bees) to identify locations richest in resources, communicate their location to comrades, exploit them quickly, and abandon depleted locations rapidly in favor of alternate sources. Honey bees manage to meet these challenges with startling efficiency by a very simple method: each forager evaluates each site they visit; if a forager is excited by the resource richness of the site, she returns and tells a subset of her comrades the location of the resource and her own relative level of excitement (via mathematical dance language). Bees who receive her signal decides whether or not she was excited enough to merit their own trip to the site. If they go, they either return just as excited to recruit others, or else disagree, decide the site was not exciting enough, and search for a new site themselves or wait for another comrade to recruit. This system fulfills many desirable features: excitement waxes and wanes endogenously with site quality, sites are exploited while also searching for new sites, individuals identify new sites that do not fit the current predominant interest, and attention accrues very rapidly at any site consensus deems worthwhile without the need for bees to agree a priori on any single definition of "exciting."

Putting Bees to Work in Cyber-Defense

In this article we define HONIED: Hive Oversight for Network Intrusion Early Warning using DIAMoND — a bee-inspired method for fully distributed cyber defense. Our research is the first to investigate the potential for a self-organizing anomaly detection system inspired by the distributed algorithms colonies of honey bees use to forage efficiently to provide appropriate, dynamic detection thresholds for anomalous event patterns on computer system networks to improve early detection and mitigation methods to counter malicious threats.

Our approach addresses some of the main challenges of distributed defense strategies. The proposed system allows for cooperation between sensors in an arbitrary virtual topology and does not rely on sharing the particulars of the underlying event, but only the pattern of "excitation" seen in the sensors. By its nature this data does not contain any individually sensitive information, or even any information about the specific attack. We expect that overcoming organizational hurdles that may prevent sharing of such data would be far easier. For the same reason, our scheme easily addresses the third and fourth challenges; because the data shared is very simple (not even individual values for detection thresholds are shared), there is no question of creating semantic equivalence or complex correlation techniques. Finally, the scheme enables sensors to self-tune their individual detection threshold values using a feedback mechanism. When new attack patterns appear, the sensors learn by cooperation to sense them — it takes some time, but there is no prior modeling that has to be applied to the sensors. That makes our scheme especially appealing for detecting novel network attacks assuming that some controls (e.g., local intrusion detection systems) are able to detect their symptoms.

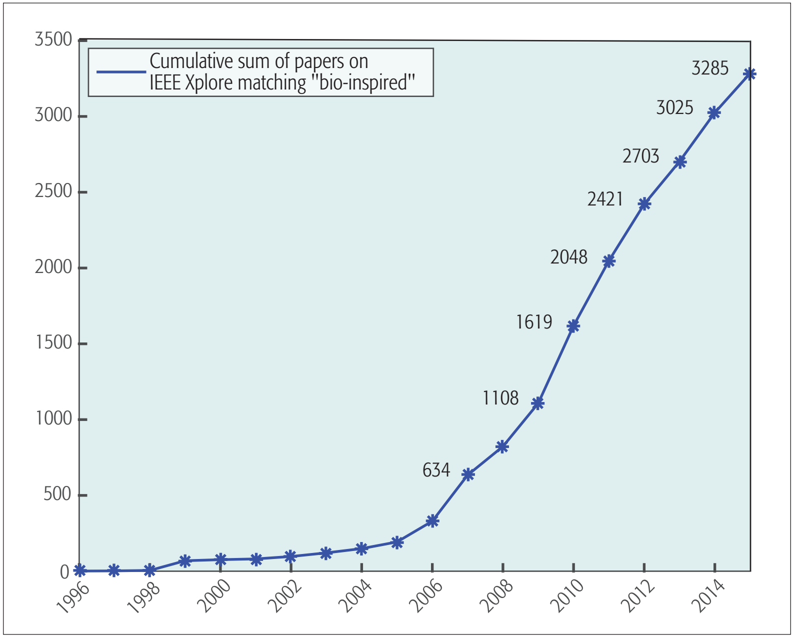


Figure 1. Literature in bio-inspired algorithms.

Related Work

There have been several proposals for fully distributed systems [3–7]. Locasto *et al.* proposed a fully distributed peer-to-peer (P2P) intrusion detection system (IDS) called Worminator [4]. The system creates and shares between the federations of nodes compact watchlists of IP addresses encoded in Bloom filters. Another P2P approach for collaborative intrusion detection is proposed by Zhou *et al.* [5]. It implements a distributed hash table (DHT) system to share detection information. Each peer submits its blacklist to a fully distributed P2P overlay. The participating nodes are notified if other peers are attacked by the same source. However, both methods use a single traffic feature, which might be too restrictive for detecting some important characteristics of large-scale intrusions.

In a distributed IDS proposed by Dash *et al.* [6], local detectors use a binary classifier to analyze incoming/outgoing host traffic and raise an alarm if a threshold value is crossed. Through their information sharing system (ISS), those alarms are sent to a random set of global detectors that generate a global view of security status of the system being monitored. DefCOM [7], which is a distributed system for DDoS mitigation, consists of three types of nodes: core, classifier, and alert generator nodes. It implements an overlay communication protocol between source, victim, and core networks to detect and block the attack at the source. One of the main drawbacks of both systems, however, is the separation of

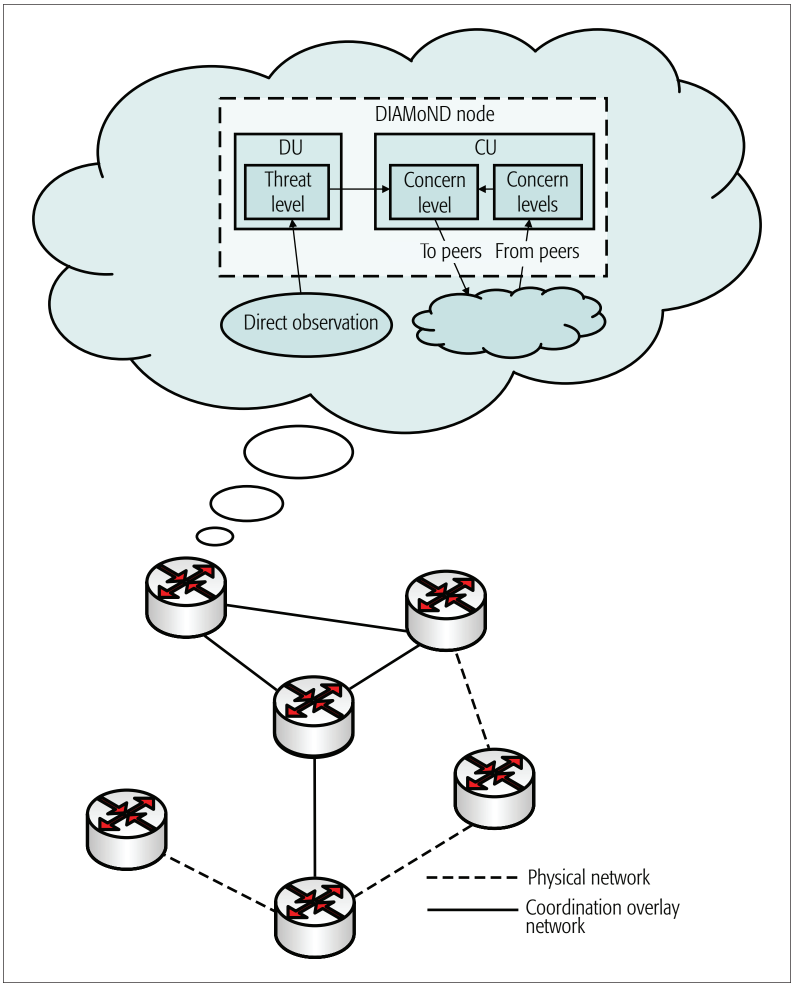


Figure 2. DIAMoND architecture.

different types of nodes and the need for the systems to coordinate messages between them.

While bio-inspired (cf. Fig. 1, e.g., [8]), and honey-bee-based algorithms in particular, are not new [9, 10], our approach is among the first to apply them to distributed-decision-driven cyber-security systems.

Honey Bee-Inspired Detection System

Forming the Analogies with Honey Bee Foraging

In honey bee foraging [2], system participants do not define the search target a priori, instead letting participants identify anomalies (resources) as they encounter them. This feature is one of the most important benefits we anticipate from adopting this bio-inspired perspective, particularly when detecting complex network attacks that might coincide with each other (in which there are no known patterns for which to look). In honey bee foragers, if enough participants identify a location as a valuable target (i.e., an anomaly), it becomes an anomaly by definition. Furthermore, as an anomaly is handled (i.e., resources are exploited), participants gradually lose interest, ceasing to identify the location as anomalous.

Another important feature of the system is that foragers who act as early scouts return to recruit additional foragers to help exploit identified anomalies (i.e., resources). They communicate not only the location, but also their "relative excitement" about the quality of the discovered resources to all other bees within range, called the foraging dance floor. This is functionally equivalent to a nonparametric description of perceived importance of the identified target, allowing very rapid and low-overhead communication and census-taking for collaborative decision making. This real-time collaborative definition of anomalies makes the system uniquely suited to discover novel targets by eliminating the need to employ any form of uniform template for comparison or recognition. We critically also adopt these features in our algorithm design.

Basing our algorithm on this system, instead of traditional distributed network anomaly detection (in which we must have a list of known patterns that indicate attacks and/or legitimate traffic), we instead allow emergent consensus to draw attention to patterns, even if some participants would not have identified the pattern as indicating an attack if assessed only independently.

System Basics

We use Distributed Intrusion/Anomaly Monitoring for Nonparametric Detection (DIAMoND): a nonparametric, fully distributed coordination framework that decouples local intrusion detection functions from network wide coordination. DIAMoND first builds coordination overlay networks on top of physical networks. DIAMoND then dynamically combines *direct observations* of traditional localized/centralized network IDS (NIDS) with knowledge exchanged with other coordinating nodes called *neighbors* to dynamically detect anomalies of underlying physical systems. Specifically, coordinating nodes in DIAMoND, analogous to honey bees, exchange generic nonparametric *levels of concern* between neighbors that reflect the observed probability of network attacks without elaborating any further details on the attacks themselves. As a result, the coordination layer of the DIAMoND framework can readily be coupled with any local detection schemes without the need for increasing the detection feature sets. The coordination network layer is also decoupled from the underlying physical network layer to facilitate flexible coordination strategies based on, for example, previously observed correlated behaviors, instead of being artificially limited to direct connectivity or geographical proximity. Interactions inside DIAMoND are limited to local neighborhood (e.g., one- or two-hop neighbors) in the overlay network, thus ensuring system scalability linear to the coordination network density instead of network size. While in general there can still be potential risks for recovery of sensitive information from the sharing of only nonparametric descriptors, in this case, since there is no need for/assumption of a uniform individual detection algorithm for local determination of level of excitement/concern across participating nodes, or even for a single node over time, no inference can be made simply from the nonparametric information shared about more sensitive features. The overall architecture of DIAMoND thus allows preservation of potentially sensitive

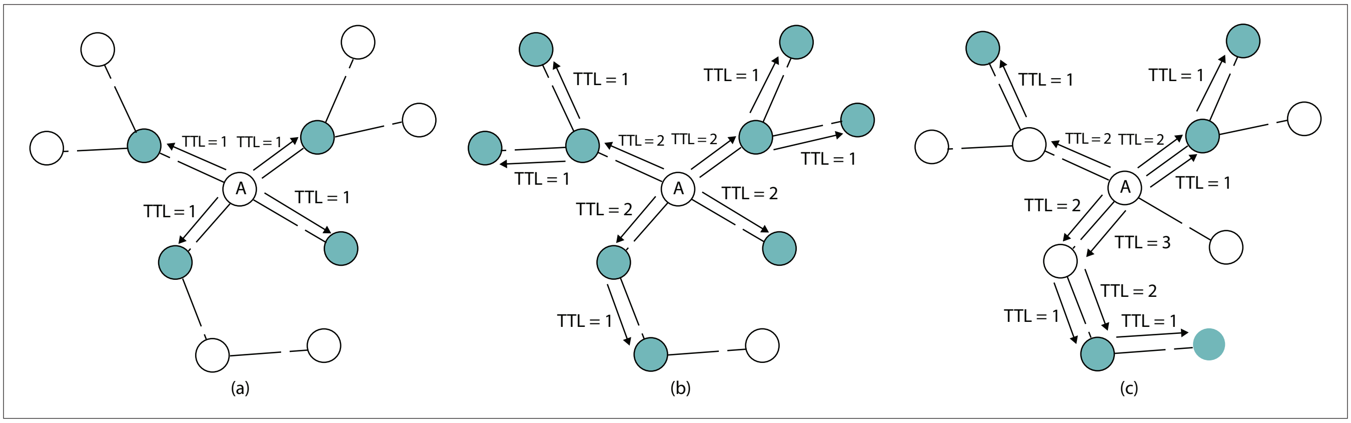


Figure 3. Neighborhood strategies: a) hop limit TTL = 1; b) hop limit TTL = 2; c) correlated attacks neighborhood.

information of individual participating parties, which eases deployment of DIAMoND across political and administrative boundaries.

System Design

Architecture Overview

DIAMoND is deployed over multiple *nodes* (switches, middleboxes) in a fully distributed architecture (Fig. 2). We define a node's *neighborhood* as a subset of all nodes with which it directly exchanges nonparametric alert-related information. Neighborhoods are dynamic and can change over time based on, for example, previously observed correlated behaviors or changes in the network topology. Two collaborating nodes enjoy a symbioti, mutual relationship, meaning both must authenticate each other and agree to join each other's neighborhoods. Furthermore, each node is equipped with two functional units: a detection unit (DU) and a coordination unit (CU). The former is responsible for the data-driven individual assessment of the so-called *threat level* — the level of likelihood that an intrusion is occurring based on the *direct observation* reported by local NIDS and/or firewall implementation. The latter calculates the *concern level*, which is a function of its own *threat level* and the *concern levels* of its neighbors (Fig. 2).

Detection Unit

Any detection or security intelligence such as NIDS or firewalls can be implemented in a DU as long as there is an appropriate plug-in to a CU to translate the output of the DU to the nonparametric threat level. Additionally, there must be an incorporated appropriate response by the DU to different levels of concerns of its neighbors (e.g., tuning of sensitivity thresholds). To foster interoperability, we do not require the extraction and provision of any potentially sensitive and/or incomparable attack details. In fact, a node may choose any local anomaly detection method independent from any other node(s), thereby making it difficult for an attacker to manipulate the local anomaly detection's influence on the CU network by making it harder to predict what types of traffic may trigger an individual, local intrusion warning. These features greatly increase the potential of such a system to be able to detect diverse characteristics of large-scale network attacks, depending on a variety of local detection algorithms adapted to DIAMoND.

Coordination Unit

Each of the participating nodes has an internal set of *sensitivity thresholds* corresponding to their "native" detection algorithms. These sensitivity thresholds are updated dynamically over time, and there is no a priori assumption of their uniformity across nodes. Since each node may employ its own local anomaly detector, these thresholds are also completely independent of each other. The sensitivity threshold is a function of the observed threat level and the level of concern of each node's neighborhood. Note that even if the sensitivity threshold is dynamic, it can be updated within a certain predefined range to prevent malicious tuning.

At each time instance, each node computes a function of the observed threat level, which is the individual data-driven assessed level of the likelihood that an anomaly is occurring.

We assign values *low*, *med*, *high* to the threat level for each node in each time instant based on the traffic observed in the local intrusion detection on that node. Values are defined such that low indicates a completely normal classification, med indicates that traffic patterns have exceeded some fixed numbers of standard deviations from normal but have not yet exceeded the rate limiting threshold to be considered an attack, and high indicates classification of a current attack by the local anomaly detector.

Each node has a level of concern at every time instant, which is a function of both the previously assessed threat level and of the total impact of the concerns of all nodes within its neighborhood computed by our naïve excitation algorithm that takes discrete values low, med, high. Values are defined such that low indicates a consensus between a node's neighbors and normal network state, med indicates that traffic patterns observed within a neighborhood have deviated from normal traffic distributions but have not yet exceeded some thresholds to be considered an attack, and high indicates classification by the node's neighborhood of a current attack.

Finally, each node determines the strength of influence of the levels of concern from its neighbors. This strength allows a node to tune preference between sensitivity and specificity provided by the collaborative network. We here

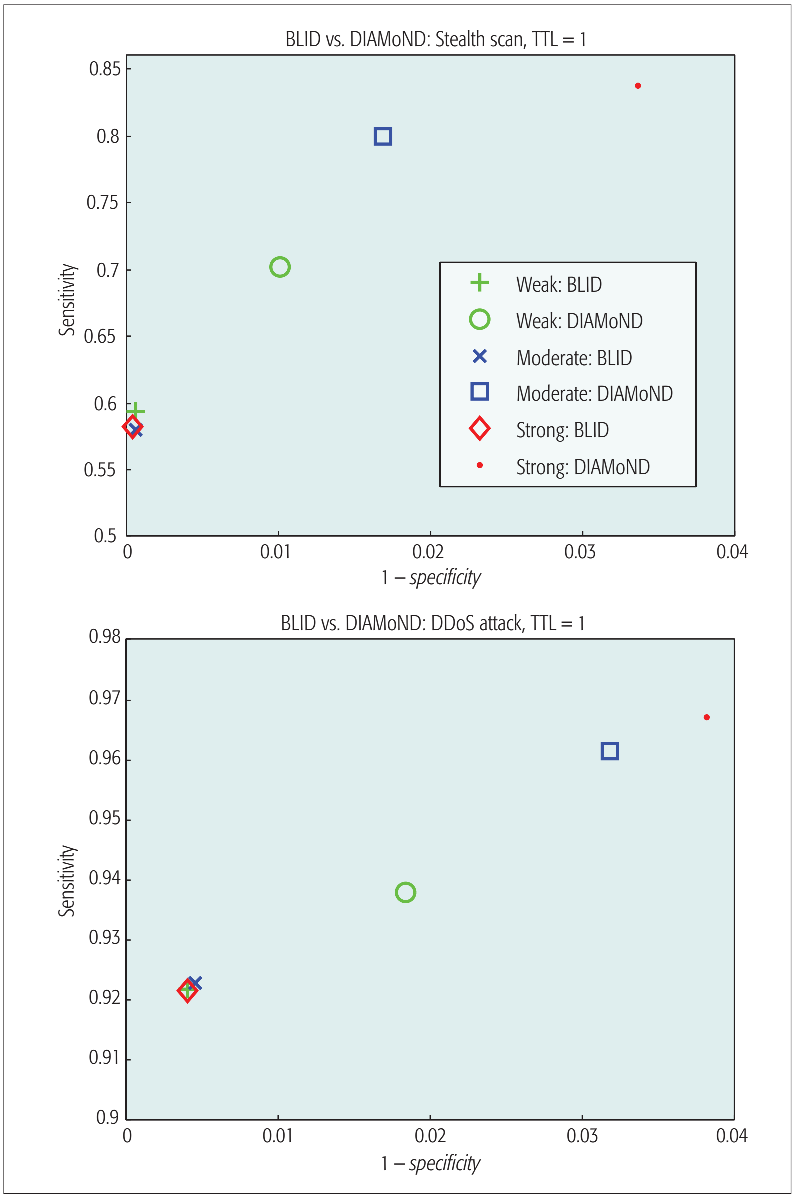


Figure 4. Comparison of DIAMoND vs. BLID for network-wide stealth scans (top) and DDoS attacks (bottom). We also explored the impact of either strengthening (strong) or weakening (weak) the influence of network neighbors to show the robustness of effect and test system sensitivity to individual-node-level detection accuracy.

present the full results for a moderate strength of influence, but results from other choices may be found in [11].

Neighborhood Strategies

Honey bees incorporate the influence of other nodes into their decision on whether or not to reinforce the signal as discussed earlier. We define and investigate two different strategies for creating the "areas" or neighborhoods to maximize the flow of meaningful information while minimizing the number of connections.

The first strategy is based on a hop limit that reflects the geographical or administrative distance between neighbors. In the simplest but very effective form, we define a neighborhood of a node by direct physical or logical connection. We also attempt to empirically verify the application of the extended neighborhoods by increasing the time to live (TTL) value (Figs. 3a and 3b). Another strategy, depicted in Fig. 3c, consists of correlating previously observed attacks and constructing neighborhoods based on the assumption that malicious activity may reoccur and be launched from the same set of compromised machines and/or against the same victims (networks, servers).

Evaluation Testbed

We have developed our prototype communication protocol as an OpenFlow controller in the POX environment[3] and evaluated it using the Mininet 2.0 network emulator.[4] Our initial software system deployment consists of 20 nodes due to computational constraints and up to 20 end-user machines connected to each node. The full specification together with the communication protocol is available to the public.[5]

In this article, we test the performance of the algorithm on an "extended star" physical topology that represents a tree of 19 links which is generated by initiating the graph with a "root" node and then attaching each subsequently created node to one of the already existing nodes in a uniform fashion.

In our experimental evaluation, we use traffic captured from the trans-Pacific line (samplepoint-F, 150 Mb/s).[6] The traffic is labeled by the MAWI working group as anomalous or normal using an advanced graph-based method that combines responses from independent anomaly detectors built on principal component analysis (PCA), the gamma distribution, the Kullback-Leibler divergence, and the Hough transform [12]. Then we develop our method based on an *X-means* algorithm. Finally, we filter all traffic labeled as anomalous by each classification method and use the remaining traffic in our benchmark traffic generator.

Each node has been equipped with a sampling detection algorithm for detecting SYN flooding attacks and TCP portscan activity [13]. The method considers TCP connections as legitimate if it samples one of multiple acknowledgment (ACK) segments (with disabled SYN flag) coming from the server. It defines two traffic features: a number of outgoing SYN segments to corresponding incoming ACK segments per source and per destination IP address. The method is combined with a rate limiting scheme — if the traffic rate is less than or equal to a predefined rate for a given IP address, it is allowed to pass the filter, whereas traffic that exceeds the rate is dropped. For the purpose of this study, we refer to the above-described algorithm as benchmark local intrusion detector (BLID). To meet the needs of our system, we extend the proposed algorithm and define the range of sensitivity rate limiting thresholds as well as the plug-in that translates the output of the algorithm to the nonparametric thread level.

We evaluate the capability of our system using two predominant attacks exploiting TCP protocol: network-wide SYN stealth scans and SYN flooding attacks that are launched from a selected percentage of the network nodes, which are considered compromised and take part in a coordinated distributed attack. For more details

[3] https://openflow.stanford.edu

[4] http://mininet.org

[5] http://mkorczynski.com/diamond.html

[6] http://mawi.wide.ad.jp/mawi

Sensitivity		1 – *specificity*		Accuracy		
BLID	DIAMoND	BLID	DIAMoND	BLID	DIAMoND	Gain
Stealth scan, TTL = 1 neighborhood						
0.58 (±0.02)	0.8 (±0.015)	6.2e^{-4} (±1,5e^{-4})	0.017 (±0.003)	0.889	0.935	0.047
Stealth scan, TTL = 2 neighborhood						
0.557 (±0.021)	0.787 (±0.021)	7.5e^{-4} (±5.2e^{-4})	0.019 (±0.003)	0.889	0.932	0.045
Stealth scan, TTL = 3 neighborhood						
0.568 (±0.029)	0.793 (±0.029)	6.1e^{-4}(±1.7e^{-4})	0.02 (±0.003)	0.887	0.932	0.045
Stealth scan, attack correlation neighborhood						
0.528 (±0.027)	0.752 (±0.027)	5.55e^{-4} (±1.3e^{-4})	0.02 (±0.003)	0.891	0.931	0.041
DDoS attack, TTL = 1 neighborhood						
0.923 (±0.012)	0.962 (±0.01)	0.005 (±7.3e^{-4})	0.032 (±0.004)	0.95	0.964	0.014

Table 1. Sensitivity, 1 – *specificity*, accuracy of BLID and DIAMoND, and the accuracy gain of DIAMoND over BLID. Performance at low TTL demonstrates significant benefit without increased communication overhead costs associated with higher TTLs.

Sensitivity measures the proportion of malicious packets that are correctly identified as such, specificity measures the proportion of legitimate packets that are correctly identified as such, whereas accuracy measures the proportion of correctly identified malicious and legitimate to all the packets.

on the tesbed, we refer the reader to our previous work [11].

Emulation Results

Criteria of Detection Evaluation

To assess the performance of DIAMoND, we consider three meaningful metrics: sensitivity, specificity, and overall system accuracy. Sensitivity measures the proportion of malicious packets that are correctly identified as such, and specificity measures the proportion of legitimate packets that are correctly identified as such, whereas accuracy measures the proportion of packets correctly identified malicious and legitimate to all the packets.

Also, we quantify the additional information that is gained by deploying our system on top of BLIDs. In other words, we ask by how much, if at all, the inclusion of the DIAMoND collaboration among nodes improves their accuracy relative to their use of only the local detection algorithms in isolation. In order to evaluate the information gain we use an information theoretic approach, Kullback-Leibler (K-L) divergence.

It is important to recall that the potential for improvement in accuracy is scaled by the percent of malicious packets. Since in the case of network-wide stealth scans malicious packets constitute a smaller percentage of all network traffic, the increase in accuracy is strictly bounded, meaning that, for example, 0.045 represents a substantial improvement relative to the range possible for improvement.

Detection Performance

Figure 4 shows a *sensitivit*y as a function of 1 – *specificity* for network-wide stealth scans (top) and DDoS attacks (bottom) in an overlay network where neighborhoods are created on the basis of direct physical connections (TTL = 1). We present results that reflect participating nodes assigning a moderate level of influence from the concern levels of their neighbors to their own decision, but then also present results from both weakening and strengthening that influence for comparison. The results for stealth scans indicate that the more influence nodes assign to their neighbors' concern, the greater their improvement in sensitivity, without compromising specificity in comparison to BLID systems operating independently. The fact that 1 – *specificity* does not exceed 3.5 percent (in the worst case) comes from two reasons:

- Precise calibration of the rate limiting sensitivity thresholds. For example, the consensus of *level of concerns* of neighbors cannot reduce the sensitivity threshold of a chosen node below some pre-calibrated minimal value.
- The *level of concern* of a node signals the anomaly, while the decision about the assigning particular flows to *legitimate* or *malicious* classes remains with the DU.

As with the sensitivity improvements, the overall information gain of DIAMoND calculated over the accuracy of BLID increases as participating nodes increase the influence of the input from their neighbors (approximately twice as large for *moderate* and *strong* as for *weak*; Table 1).

In the evaluated attack scenarios, we observe no major distinction in the detection accuracy and information gain regardless of the neighborhood strategies (Table 1).

Finally, our results show less significant improvement in sensitivity of our system over BLID systems operating independently for DDoS attacks: between 1.6 and 4.5 percent (Fig. 4 and Table 1). We also observe that the information gain of the overlay detection system is lower (although always positive) in comparison with low-rate malicious activity, but the system can react close to the source of the attack more effectively and thereby reduce the collateral damage to a minimum.

Minimal and Marginal Deployment Gain

Deployment of networked services across administrative boundaries usually has to take place progressively. In this section, we try to understand the minimal deployment percentage

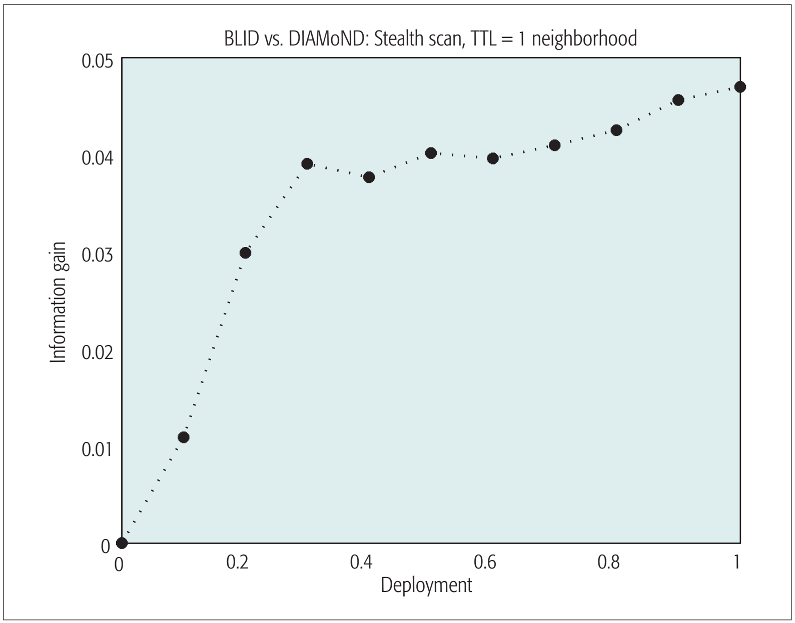


Figure 5. Information gain of DIAMoND over BLID.

needed for DIAMoND to have significant performance impact and marginal performance gain with additional deployment.

To quantitatively evaluate deployment gain, we adapt a calculation of "offline marginal utility," originally proposed to analyze the impact of additional metrics, to instead compute the incremental information gain for each additional node (relative to the information achieved with BLID). We refer the reader to some relevant literature for more details [11].

Figure 5 provides an example analysis of the deployment gain for a 20-node network under network-wide port scan probing. This figure shows a point of diminishing return such that, after 30 percent of the nodes participate in DIAMoND, the information gain is close to that achieved when all nodes are participating, and the marginal deployment gain from increasing participation is insignificant. On the other side, even when there are only 10 percent nodes participating, the information gain is already over 0.01. When 20 percent nodes are participating, the information gain reached a significant 0.03. We thus concluded that, in this case:

- Minimal effective deployment is 10 percent of the network nodes participating.
- Marginal gain is maximized at 20 percent deployment.

DIAMoND plateaus after 30 percent deployment, with minimal value gained by having additional nodes participating.

As our immediate next step we plan to explore the scalability of DIAMoND coordination protocol, and apply it to a broad set of deployment scenarios and real-network topologies.

Conclusions

In this article we investigate the potential for a self-organizing, nonparametric distributed coordination framework inspired by those observed naturally in colonies of honey bees to provide dynamic individual detection thresholds for anomalous event pattern detection on networks.

To illustrate its application, we couple DIAMoND with local anomaly detection schemes for network-wide stealthy port scan and SYN-flooding-based DDoS and evaluate its performance on an emulation testbed. DIAMoND demonstrated up to 20 percent enhancement in sensitivity without sacrificing specificity. In this article, we also systematically investigate several automated coordination neighborhood construction strategies and find that DIAMoND exhibits stable performance gain over different neighborhood strategies. This leads us to conclude that DIAMoND is robust to neighborhood size. Deployment impact shows that DIAMoND quickly reaches an information gain plateau after 30 percent of network nodes participate in coordination, which enhances the deployability of DIAMoND. It allows multiple entities, which may be functionally and/or legally prohibited from sharing cyber data, to leverage each other's insight and increase their effectiveness in cyber defense. Furthermore, DIAMoND enables real-time adaptation, eliminating the identification-designed-response delay inherent in defenses that react to known and predefined threats, and allowing active defense for emerging novel network attacks.

Acknowledgments

The U.S. Department of Homeland Security sponsored this research under the Air Force Research Laboratory (AFRL) agreement number FA8750-12-2-0232. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes. This work represents the views of the authors and does not represent official policies or endorsements, either expressed or implied, of AFRL or the U.S. Government.

References

[1] A. Noroozian *et al.*, "Developing Security Reputation Metrics for Hosting Providers," *Proc. USENIX CSET*, 2015, pp. 1–8.
[2] M. L. Winston, *The Biology of the Honey Bee*, Harvard Univ. Press, 1991.
[3] C. V. Zhou, C. Leckie, and S. Karunasekera, "A Survey of Coordinated Attacks and Collaborative Intrusion Detection," *Computers & Security*, vol. 29, no. 1, 2010, pp. 124–40.
[4] M. Locasto *et al.*, "Towards Collaborative Security and P2P Intrusion Detection," *Proc. IEEE IAW*, 2005, pp. 333–39.
[5] C. V. Zhou, S. Karunasekera, and C. Leckie, "A Peer-to-Peer Collaborative Intrusion Detection System," *Proc. IEEE ICON*, vol. 1, 2005.
[6] D. Dash *et al.*, "When Gossip Is Good: Distributed Probabilistic Inference for Detection of Slow Network Intrusions," *Proc. Nat'l. Conf. AI*, vol. 2. AAAI Press, 2006, pp. 1115–22.
[7] M. Robinson *et al.*, "DefCOM: Defensive Cooperative Overlay Mesh," *Proc. DARPA Info. Survivability Conf. and Expo.*, vol. 2, 2003, pp. 101–02.
[8] W. Mazurczyk and E. Rzeszutko, "Security — A Perpetual War: Lessons from Nature," *IT Professional*, vol. 17, no. 1, 2015, pp. 16–22.
[9] D. Karaboga and B. Akay, "A Survey: Algorithms Simulating Bee Swarm Intelligence," *Artificial Intelligence Review*, vol. 31, no. 1–4, 2009, pp. 61–85.
[10] G. A. Fink *et al.*, "Defense on the Move: Ant-Based Cyber Defense," *IEEE S&P*, vol. 12, no. 2, 2014, pp. 36–43.
[11] M. Korczyński *et al.*, "DIAMoND: Distributed Intrusion/Anomaly Monitoring for Nonparametric Detection," *Proc. IEEE ICCCN*, 2015, pp. 1–8.
[12] R. Fontugne *et al.*, "MAWILab: Combining Diverse Anomaly Detectors for Automated Anomaly Labeling and Performance Benchmarking," *Proc. ACM CoNEXT*, 2010, pp 1–12.
[13] M. Korczyński, L. Janowski, and A. Duda, "An Accurate Sampling Scheme for Detecting SYN Flooding Attacks and Portscans," *Proc. IEEE ICC*, 2011, pp. 1–5.

Biographies

Maciej Korczyński (maciej.korczynski@tudelft.nl) is a postdoctoral scientist in the cybersecurity research group at Delft University of Technology. He received his Ph.D. degree in computer science from Grenoble University of Technology, France, in 2012. Previously, he was a postdoctoral research associate at Rutgers University, New Jersey (2013–2014). His research interests include encrypted traffic classification, security of the TLS and DNS protocols,

passive and active Internet security measurements, incident data analysis, economics of cybersecurity, anomaly and attack detection, and bio-inspired cybersecurity.

ALI HAMIEH (adhamieh@gmail.com) is a research scientist at Rutgers University. He received a Ph.D. degree from the University of Versailles in France for his thesis, *Security in Wireless Ad Hoc Networks: The Cases of Jamming Attacks and Greedy Behaviors*. His research interests include network design and security.

JUN HO HUH (junho.huh@honeywell.com) is a research scientist at Honeywell ACS Labs. He received his Ph.D. in the field of cybersecurity and trustworthy computing from the University of Oxford. Since joining Honeywell, he has been involved in numerous intrusion detection projects for embedded systems, developing specification- and outlier analysis-based intrusion detection sensors for smart meters and flight controllers. His research interests also include designing intuitive cybersecurity dashboards and usable authentication solutions for control systems.

HENRIK HOLM [M] (henrik@forestglenresearch.com) is an independent consultant with Forest Glen Research, LLC. He received his Ph.D. from NTNU, Norway, in 2002. He previously worked as a postdoctoral researcher and lecturer with the University of Minnesota, and at Honeywell ACS Labs. His current interests include medical and embedded device security, signal processing, and machine learning.

S. RAJ RAJAGOPALAN (siva.rajagopalan@honeywell.com) is a research scientist at Honeywell ACS Labs. He received a Ph.D. in the field of theoretical computer science from Boston University. Since joining Honeywell in 2011, he has been leading the effort at ACS Labs of incorporating the fruits of the latest cybersecurity research into the vast portfolio of control systems in Honeywell ACS. His research interests also include using techniques from socio-cultural anthropology to address cybersecurity problems, and the study of the interactions between security and safety in modern buildings.

NINA H. FEFFERMAN (feffermn@dimacs.rutgers.edu) is an associate professor in both DEENR and DIMACS at Rutgers University. She received her Ph.D. in biology from Tufts University in 2004, and her M.S. and A.B. in mathematics from Rutgers University in 2001 and Princeton University in 1999, respectively. Her research explores evolutionary biology, epidemiology (in humans and wildlife), cybersecurity, and any other complex systems where the success of individuals involves the success of the group to which they belong.

Bio-Inspired Cybersecurity for Wireless Sensor Networks

Salim Bitam, Sherali Zeadally, and Abdelhamid Mellouk

The authors present a careful review of different bio-inspired techniques developed for improving cybersecurity of CPS using WSNs. Additionally, they propose a generic bio-inspired model called Swarm Intelligence for WSN Cybersecurity (SIWC) that addresses drawbacks of prior bio-inspired approaches.

Abstract

Rapid advances in information and communication technologies have led to the emergence of cyber-physical systems (CPSs). Wireless sensor networks (WSNs) play a pivotal role in CPSs, particularly for operations such as surveillance and monitoring. However, these WSNs are subject to various types of cyberattacks that can cause damage, theft, or destruction of sensitive data, in addition to disruption of services provided by CPSs. To strengthen cybersecurity in WSN-enabled CPSs, various researchers have proposed a new category of efficient algorithms, inspired by biological phenomena. We present a careful review of different bio-inspired techniques developed for improving cybersecurity of CPSs using WSNs. Additionally, we propose a generic bio-inspired model called Swarm Intelligence for WSN Cybersecurity (SIWC) that addresses drawbacks of prior bio-inspired approaches.

Introduction

The widespread deployment of information technology (IT) in various cyber-physical systems (CPSs) such as smart grids, healthcare platforms, and computer networks has made them vulnerable to various types of security attacks known as cyberattacks. Such attacks are becoming increasingly sophisticated and dangerous, attempting to gain unauthorized access to a service or data, or trying to compromise a computational system's confidentiality, availability, or integrity. The last few years have brought a tremendous increase in the number of cyberattacks, along with the emergence of various types of cybercriminals who constantly develop new attack techniques.

According to the Ponemon Institute [1], the average consolidated total cost of a data breach, based on a recent 2015 study of 350 companies spanning 11 countries, is $3.8 million worldwide, up from $3.5 million a year ago. The same study found that the cost of a data breach is $154 per stolen record containing sensitive information, up from $145 in 2014. Due to the increasing cost of cyberattacks and our heavy reliance on computer systems and technologies, cybersecurity has emerged as an important research field to control and prevent such access.

Trends in Cybersecurity with Wireless Sensor Networks (WSNs)

CPSs adopt and deploy technologies extensively, such as wireless sensor networks (WSNs) for many application domains. In particular, WSNs contribute to ensure the cybersecurity of CPSs, where sensors may dynamically collect physical information through a cooperative process that helps detect and mitigate potential future cyberattacks (as shown in Fig. 1).

In the literature, WSNs have been heavily used to support various surveillance and security functions. For example, sensor networks have been deployed to support surveillance capabilities such as threat-presence detection within security-sensitive and hostile regions such as a militarized area, border protection, etc. To support surveillance, a WSN has been proposed in [2] to detect and determine the direction of movement of intruding personnel and vehicles (i.e. target tracking) in the sensitive zone. For many of the surveillance functions, deployed sensors cooperate with each other to detect an imminent approaching mobile threat and are able to self-organize to provide a relevant, timely, and concise net-centric view of the surveillance field. This information helps to enhance decision-making abilities for command and control, intelligence, surveillance, and reconnaissance tactical mission planning. To enhance these decision-making abilities, sensor nodes should be able to forward threat information to a gateway node called the sink node. In such an environment, sensor nodes may be compromised by intruders to disrupt sensed and transmitted data by injecting false data reports.

Medical monitoring can also be cited as a healthcare service provided by wearable and implantable body sensors connected in the well-known body sensor network (BSN). A typical BSN is composed of a number of miniature, lightweight, low-power sensing devices and wireless transceivers. These sensor networks are used to capture large amounts of data containing information about the patient's health status, which is then stored in some database. Health status data commonly includes information such as blood pressure, heart rate, distance traveled through walking/running, playing activ-

Salim Bitam is with the University of Biskra, Algeria; Sherali Zeadally is with the University of Kentucky; Abdelhamid Mellouk is with the University of Paris-Est.

0163-6804/16/$25.00 © 2016 IEEE

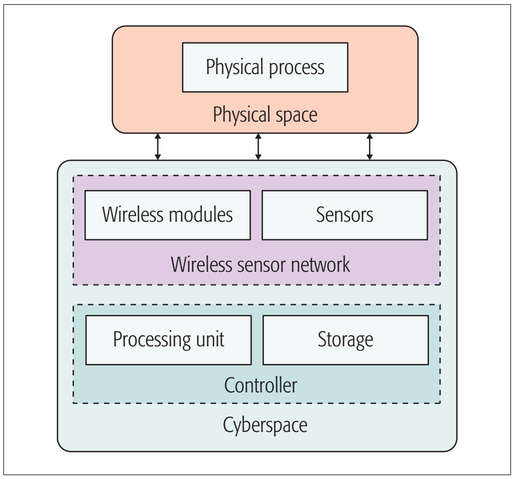


Figure 1. WSN-enabled cyber-physical systems.

ities, and surroundings (e.g. room temperature). This collected information helps the early detection of emergency conditions, diseases in at-risk patients, as well as the monitoring of chronic disease, elderly people, postoperative rehabilitation patients, and persons with special disabilities [4]. However, in this case a BSN may be exposed to a malicious party who can exploit various serious security threats to compromise the healthcare service and prevent patients from reaching available healthcare facilities.

Earlier discussions about the application domains (surveillance of sensitive regions and healthcare) demonstrate that WSNs are critical to provide vital solutions and cybersecurity that protect countries', organizations', and individuals' digital equipment resources and services from unintended or unauthorized access/change, disruption, or destruction. Attacks on WSNs can potentially affect the infrastructure's/application's entire operation (along with data confidentiality, integrity, and availability) given the integral role they currently play.

Motivations for Using Bio-Inspired Approaches for Cybersecurity

In the literature, several traditional approaches were proposed [3] to cope with WSNs' cybersecurity. However, these traditional security algorithms are not very effective for WSNs for the following reasons:

- In previously proposed traditional security approaches, compromised nodes are hard to detect because these algorithms typically consider one centralized node (i.e. a base station) responsible for detecting cyberattacks for large-scale networks. This situation leads to a heavyweight security process performed by the base station, which could potentially miss several attacks.
- Conventional security solutions do not scale well with the rapid increase in information or processing required by the massive influx of large amounts of data. Furthermore, the computational complexity of cyberattacks requires security solutions that are more scalable, robust, and flexible than traditional security methods can offer [5].

Therefore, recently a new category of methods has emerged that is inspired by biological phenomena such as biological evolution, biological immune system, and swarm intelligence. Bio-inspired techniques for WSN cybersecurity are initially motivated by the successful adaptive defense process of insects, such as ants against threats where they can ramp up their defense rapidly, and then resume routine behavior quickly after an intruder has been stopped. Bio-inspired approaches are highly scalable, use lightweight architectures, and are less resource-constrained compared to traditional security solutions.

Bio-inspired techniques for WSN cybersecurity are initially motivated by the successful adaptive defense process of insects, such as ants against threats where they can ramp up their defense rapidly, and then resume routine behavior quickly after an intruder has been stopped.

Bio-Inspired Methods and Their Application to Cybersecurity

Here, we briefly review three bio-inspired approaches aimed at improving cybersecurity (although not specifically for WSNs).

Genetic Algorithm (GA)

GA was proposed in [7] to create a moving target defense where the computer configurations (operating system and/or applications) are directly manipulated to find diverse, secure configurations that are placed in service at varying periods of time. The motivation behind this idea is that alternative configurations found by GA can disrupt the attacker's knowledge about the system. Therefore, the attacker acts on false or constantly changing information that may expend more resources, thereby increasing the risk of detection. This study encodes computer system configurations as chromosomes, and the security associated with each configuration is considered as its fitness. A series of selection, crossover, and mutation processes are performed to discover secure configurations. The fitness is decayed based on the period of time it was made active. Hence, less recently used configurations will be considered less secure, which will potentially pave the way for newly discovered configurations.

Ant Colony Optimization (ACO)

The authors of [8] proposed an ant-based model (called AraTRM) to address the problem of trust and reputation management and ensure the security of data forwarding in networks. In contrast to traditional approaches, the ant-based model is considered an effective way to detect an adversary node that exists in the selected path leading to the service provider. As in nature, when ants move, they leave a secreted pheromone substance to inform their nest mates of possible food discovery. Similarly, if the consumer is satisfied, he increases a score from source to destination (i.e. pheromone) on the global path from its location to the service provider, thereby rewarding this path as secured so that other consumers might use this path. On the other hand, pheromone values along the path to the malicious service provider will be punished by a victim client. As a result, the malicious service provider would be less likely to be selected as the next nodes by other consumers, because the incremented digital value (i.e. pheromone) on the edges linked to the malicious service provider is relatively small.

The DAF utilizes lightweight agents that use stigmergic (pheromone-based) communications to create useful emergent colony behaviors that ensure the protected enclave's security. Application of the DAF to energy delivery systems incorporates data from both information technology and energy delivery systems.

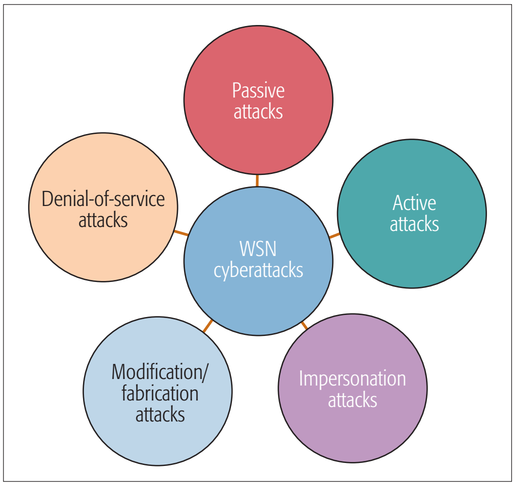


Figure 2. The five types of well-known WSN cyberattacks.

Artificial Immune System (AIS)

The authors of [9] used artificial immune system (AIS) modeling to study distributed system security issues such as identification, access level, authentication, and authorization in grid computing infrastructures. These grid systems aim to provide a safe and secure environment for anyone using recorded logs (i.e. all operations that occurred in the grid system). These logs, though, cannot ensure secure communications within the grids. To deal with this deficiency, four different groups of AIS agents were proposed: presenter, helper, memory, and killer agents. The presenter agent plays the role of an antigen that moves randomly between the network's nodes and is responsible for finding (auditing) faults, failures, or defected nodes caused by a cyberattack. If a defect occurs, the presenter agent solicits a helper agent to determine a specific killer agent for eliminating the defect's causes. The elimination operation follows a learning pattern generated and communicated to the killer by the memory agent, playing the role of lymphocytes that are known as natural killer cells (a type of white blood cells).

Next we present various WSN cyberattacks, followed by a discussion of different cybersecurity bio-inspired methods that can be applied to cope with these attacks.

WSN Cyberattacks

There are five types of well-known cyberattacks against a WSN (as shown in Fig. 2).

Passive Attack (or Eavesdropping Attack): Here, an attacker compromises and intercepts an aggregator node in the network, inspects it, listens, and reads useful data in it, trying to learn which nodes have more value within the topology (e.g. sink node or base station). Under the attacker's control, the new compromised node can be used to launch new malicious attacks. To protect nodes, WSNs should be able to conceal messages from unauthorized access (**confidentiality**).

Active Attack: In this scenario, an attacker intends to disrupt the network's functionality (**availability**). The active attacks jam communications by making changes to data already stored in the WSN in addition to modifying configuration parameters of the WSN's components (i.e. sensors). Thus, the sensors become unavailable and the expected WSN services are suspended.

Impersonation (or Sybil) Attack: In this attack, an adversary can directly replace a node's media access control (MAC) or IP address. The victim node is masqueraded as another node, which receives false data packets and compromises the trustworthiness of the information relayed.

Modification/Fabrication Attack: This attack involves unauthorized access to the WSN to modify the transmitted data packets or generate bogus data (affecting the **integrity** of the information) that is forwarded to the network nodes and to the base station. A sinkhole (blackhole) attack is a type of modification/fabrication attack that occurs when a malicious node attracts data packets sent by a sensor to a base station. A wormhole attack is another modification/fabrication attack in which the attacker records the packets at one location in the network and tunnels them to another location.

Denial-of-Service (DoS) Attack: This involves stopping the aggregation and forwarding of data in the network produced by the unintentional failure of nodes or as a result of malicious actions. DoS attacks prevent the base station from getting information from several sensors and nodes in the network.

Any of the aforementioned attacks that can potentially disrupt or destroy a network, or diminish a network's capability to provide a service, are considered a DoS attack.

Bio-Inspired Cybersecurity Solutions for WSNs

Now we review bio-inspired methods proposed for WSN cybersecurity. We classify these methods according to their biological inspiration.

Ant Colony Systems for WSN Cybersecurity

McKinnon *et al.* [6] proposed a complex-adaptive control system as a scalable approach inspired by ants' communication behavior to deal with the security risks associated with the large-scale deployments of smart grids. As with any complex system, this approach is based on inter-agent communication and the collective application of simple rules. Similar to ants, these lightweight agents move in the network and communicate using digital pheromones to alert each other about possible cybersecurity attacks. Both communication and coordination are local and decentralized, thereby allowing the framework to scale across the large number of devices that exist in the smart grid environment. This solution employs the digital ants framework (DAF), which applies lessons learned from ant foraging behaviors to address distributed cybersecurity problems. The DAF utilizes lightweight agents that use stigmergic (pheromone-based) communications to create useful emergent colony behaviors that ensure the protected enclave's security. Application of the DAF to energy delivery systems incorporates data from both information technology and energy delivery systems.

Marmol and Perez [10] proposed a bio-inspired trust and reputation model called Bio-inspired Trust and Reputation Model for WSNs (BTRM-WSN). This proposal is based on ant colony systems that select the most trustworthy node through the most reputable path, offering a certain WSN service. In particular, ants build paths fulfilling certain conditions in a graph. These ants leave some pheromone traces that help the next ants find and follow those routes.

Evolutionary Algorithms for WSN Cybersecurity

There are a few studies on WSN cybersecurity that use evolutionary algorithms because of their tradeoffs between required functional (such as accuracy) and non-functional (such as power usage and bandwidth) properties of any security solution. One such study is the reduced-complexity genetic algorithm for intrusion detection in sensor networks [11]. In this study, the authors proposed an evaluation process for sensor node attributes by measuring the perceived threat and its suitability to host the local monitoring node (LMN) that acts as a trusted proxy agent for the sink and is capable of securely monitoring its neighbors. These security attributes, in conjunction with a genetic algorithm, optimize the placement of LMNs by dynamically evaluating a node's fitness based on network integrity, residual battery power, and coverage. As a supervisor node, the sink is responsible for detecting various network misbehaviors, determining the list of sensor nodes affected, and then estimating the attack regions. This research effort [11] has shown a rapid detection of compromised nodes (with an improvement of almost 50 percent) due to the optimal placement of LMNs. Moreover, the accurate analysis of perceived threats by the sink has led to a substantial reduction in false positives and false negatives. Nevertheless, the major drawback of this genetic algorithm is its exponential computation cost for large-scale deployment of WSNs.

Particle Swarm Optimization for WSN Cybersecurity

In [12] the authors proposed a Secure Reputation Update Target Localization (SRUTL) algorithm based on PSO that addresses malicious node attacks such as Sybil attacks using target localization. SRUTL uses three phases. First, the sensor network is constructed following a uniform distribution where nodes are placed in a regular manner in the studied area, and then a stability factor (an inter-device noise) is verified at each node. This prevents malicious attacks at the node level; these nodes are considered as cluster members. A local voting scheme is performed at each node's neighborhood to elect cluster decoders (CDs) after verifying its identity. The CD with the highest reputation is selected as the cluster head (CH). Finally, a PSO algorithm is run at the CH level to detect malicious nodes in the cluster. To do that, the CH estimates the target's location and then sends an update packet to the cluster members. Once the update packet is received, each node (including the CH) computes its reputation based on its contribution for the target's location estimation and increases a local score using the signal strength at every node, as well as environmental and inter-device noise. During the next packet forwarding, the CH verifies its score with those of its cluster members. If a mismatch is found, the CH deduces that this member is a malicious node that should be ignored in any further reputation computation in the network. We note that this scheme did not address the special cases of nodes failing at the CH level, which could affect WSN security.

Artificial Immune Systems for WSN Cybersecurity

An artificial immune system called a cooperative-based fuzzy artificial immune system (Co-FAIS) was proposed in [13] to mitigate WSN DoS attacks. It is a modular-based defense system that consists of a set of agents working together to calculate the abnormality of sensor behavior or to detect the attackers. A sniffer module adapts to the sink node to audit data by analyzing the packet contents and sending the log file to the next layer called the fuzzy misuse detector module (FMDM), responsible for detecting misused nodes. The FMDM works with a danger detector module to identify danger signals' sources. The infected sources are transmitted to the fuzzy Q-learning vaccination module (FQVM) used to identify attack behavior following a reinforcement learning capability. The cooperative decision-making module (Co-DMM) incorporates the danger detector module with the FQVM to produce optimum defense strategies. An evaluation of the proposed system has shown that it improves attack-detection accuracy and yields a successful defense-rate performance against attacks after comparisons with attack-detection techniques based on a fuzzy logic controller, a fuzzy Q-learning system, and an AIS. However, a major drawback of Co-FAIS is that it needs more training time than traditional methods.

Neural Networks for WSN Cybersecurity

To detect DoS attacks at the media access control (MAC) layer of a WSN when monitoring real-time systems, the authors of [14] focus on a neural network (NN) considered as a low storage and computational time security scheme. To detect DoS attacks caused by adversaries that flood the network with packets, thereby causing collisions, two parameters were defined: the collision rate (Rc), which is the number of collisions per second detected by a node; and the arrival rate (Rr), which is the number of ready-to-send packets per second that are successfully received by a node forwarded as MAC control packets to start a data transmission. Rc and Rr are used to detect the probability of a DoS attack. Both Rc and Rr are used as inputs to the NN, and the corresponding probability of attack is represented as the targets to the multilayer perceptron (MLP). At each node, the MLP is implemented with pre-defined weights and biases, which are obtained from a trained phase. Every period, each node passes its computed values of Rc and Rr to its MLP, which produces an output that is the calculated probability of attack at that particular node. If the MLP's output (the calculated probability of attack at that particular node) is greater than a preset threshold value STH, then the node temporarily shuts itself down, and reac-

The accurate analysis of perceived threats by the sink has led to a substantial reduction in false positives and false negatives. Nevertheless, the major drawback of this genetic algorithm is its exponential computation cost for large-scale deployment of WSNs.

Study	Inspiration	Explicit introduction of user parameters	High complexity
[6]	Ant colony	Yes	–
[10] (BTRM-WSN)	Ant colony	Yes	–
[11] (LMN)	Genetic system	Yes	Yes
[12] (PSO-BASED)	Swarm intelligence	Yes	Yes
[13] (CO-FAIS)	Immune system	–	Yes
[14]	Neural network	–	Yes

Table 1. Comparison between bio-inspired methods for WSN cybersecurity.

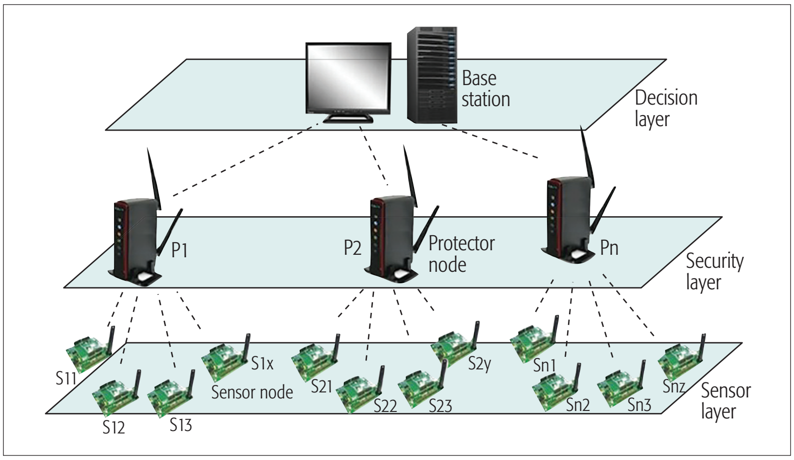


Figure 3. Our swarm intelligence for WSN cybersecurity (SIWC) model.

tivates subsequently when the attack is over. The choice of STH value is chosen depending on the extent of traffic variation in the normal operation of the WSN. Despite the good results given by this scheme in terms of the power saving due to shutting down the attacked nodes, this security mechanism might trigger a false alarm and cause the node to shut down in normal conditions without any attacks.

Discussion of Bio-Inspired Methods for WSN Cybersecurity

As Table 1 shows, the majority of the bio-inspired methods proposed for WSN cybersecurity require that the user input explicitly different parameters and variables (such as the population size for GA or pheromone value for ant systems) used by the method. In this case, because the user can choose inappropriate values, an automated process for choosing the parameter values is preferred. Moreover, some of these bio-inspired methods are of high computational complexity to be executed, especially in the case of a dense WSN. To address these major drawbacks, we propose a machine learning-based model to determine the optimal parameters for detecting more attacks without user involvement. To reduce computational complexity, this model is trained with a swarm intelligence algorithm, as we explain next.

Our Proposed Model: SIWC

We propose here a unified cybersecurity model called Swarm Intelligence for WSN Cybersecurity (SIWC). Considered as a machine learning-based approach, SIWC is an NN system trained by a swarm intelligence algorithm such as a genetic algorithm, ant colony system, particle swarm system, or some other algorithm. In contrast to traditional neural networks where general rules are given explicitly as learning rules to discover a prospective attack, SIWC is trained automatically by the swarm intelligence algorithm without requiring an explicit training process. Specifically, based on its interaction with its dynamic environment (i.e. several sources of information about prior cyberattacks), the swarm intelligence trainer can discover, formulate, and inform the NN about certain future cyberattacks. Various types of cyberattacks could be detected including active attack, passive attack, sybil attack, or DoS attack, where each one is expressed in a specific objective function established by swarm intelligence. Moreover, our approach makes use of an automatic lightweight security process (with reduced computation complexity and resource requirements) that is executed by cooperative agents. It is worth noting that the combination of swarm intelligence with neural networks has been used in the literature to solve other problems in different contexts such as data classification [15]. The architecture and functions of SIWC are explained in the following section.

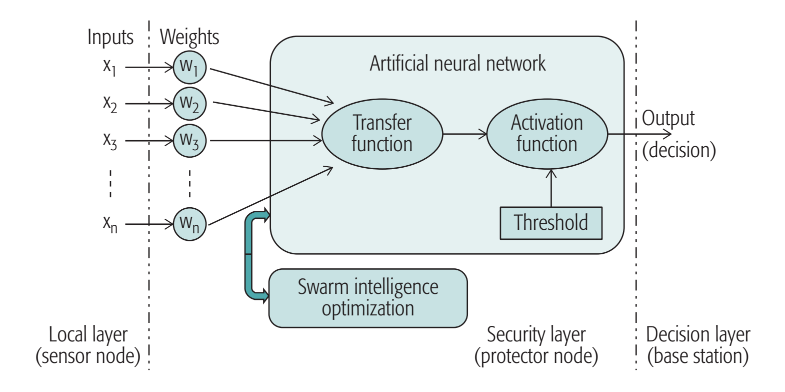
SIWC Architecture

As Fig. 3 shows, SIWC consists of three layers: the sensor layer, the security layer, and the decision layer. The sensor layer is formed by different sensors (s_{11}, …, s_{1x}, s_{21}, …, s_{2y}, …, s_{n1}, …, s_{nz}) that are responsible for sensing and collecting data transmitted to the security layer (i.e. the second layer). The security layer is a set of protector nodes (p_1, p_2, …, p_n) responsible for detecting abnormal security behaviors and attacks of a cluster of sensors. The protector node represents the head of a cluster of sensors previously formed. For example, the protector node p_n is responsible for detecting misbehaviors of its cluster composed by sensors (s_{n1}, …, s_{nz}). In order to estimate the model parameters, the Maximum-Likelihood Estimation (MLE) approach is proposed. MLE is implemented on each protector node to ensure the cybersecurity of the protector node cluster. When a threat is detected, it is sent to the decision layer, which is the WSN base station, in order to take a relevant decision, such as mitigating the attack or switching off the attacked node.

SIWC Functionality

As mentioned earlier, each sensor node periodically calculates a set of critical parameters (x_i) (e.g. collision rate or arrival rate) in its sensing range. These critical parameters are forwarded as control packets to the corresponding protector node (the cluster head) to calculate the probability of an attack (as shown in Fig. 4). These parameters are weighted according to the importance of each parameter used to define an attack among the sum of all parameters (e.g. the weight x_i for parameter x_i).

Each protector node possesses its own MLE trained using a swarm intelligence optimization algorithm after introducing the weighted parameters received from the corresponding sensors. The MLE method is trained to find the best values, which promote the highest probability of cyberattack detection. The found probabilistic value is compared to a prefixed threshold, which represents the cyberattack risk and which is determined by the swarm intelligence approach, based on previously detected attacks. Hence, the protector node informs the base station to take the appropriate action, such as shutting down a malicious node.



Inputs
Weights
x_1
x_2
x_3
x_n
w_1
w_2
w_3
w_n
Artificial neural network
Transfer function
Activation function
Threshold
Output (decision)
Swarm intelligence optimization
Local layer (sensor node)
Security layer (protector node)
Decision layer (base station)

Figure 4. SIWC functionality.

Analytical Discussion of SIWC Model

In this section, an analytical discussion of our proposed model is given. To ensure the cybersecurity of wireless sensor networks that are often deployed to control cyber physical systems, we defined a set of requirements. We discuss below how our proposed model deals with each of these requirements to provide an efficient and secure solution.

Decentralized Authentication and Integration Mechanisms: Cybersecurity solutions should use lightweight, secure authentication mechanisms and key management schemes. Key management operations must be automated. A sensor node frequently needs to communicate with its base station to report data. As a result, secure communication protocols that preserve the integrity of the data should be used to detect any unauthorized access or modification to the data without any centralized coordination.

Our proposed SIWC model is a neural network-based technique that aims to ensure authentication and integration at a local level without any centralized administration. In fact, each cluster is independently protected by its own protector node that is responsible for its management. This protector node handles all cryptographic operations such as the generation of public and private keys, digital signatures, and cryptographic hash functions.

As for authentication, the protector node locally checks all connections among WSN nodes to detect any unauthorized intrusion. To achieve this, a maximum likelihood neural network is devoted to constantly monitor the network connections. This neural network is trained by a swarm intelligence algorithm in order to obtain sufficient learning experience to detect unusual variations when an attack occurs.

Automated Connection Lock: Cybersecurity for WSNs requires the timely processing of the transmitted messages so that they are received within a pre-defined time window, otherwise anomalous delays will be logged. Also, long network delays will cause termination of the network connection automatically after a predefined period of inactivity, as well as locking a connection session after a predefined number of consecutive invalid attempts. Based on the MLE, the proposed SIWC can estimate with high accuracy different delay parameters to detect various abnormal transmissions involving various delays, inactivity periods, etc. Indeed, the MLE gives the average value of the estimated parameter (taken from several random samples) which is theoretically exactly equal or close to the population value.

Distinction between Failure Situations and Cyberattacks: WSN devices are subject to eventual failures. In order to provide safe and secure operation, the cybersecurity system should distinguish between usual failure detection and unexpected cyberattacks by performing efficient local self-scan and self-test during the occurrence of any suspected event. To achieve this goal, the swarm intelligence algorithm defines a specific objective function that evaluates different nodes' states in order to distinguish a node under attack from a node failing or malfunctioning. Consequently, a set of parameters describing an attacked node is sent to the neural network, which can trigger periodically and automatically a test and scan routine to detect any attacked node.

Minimizing Computational Complexity of Cybersecurity Solutions: Another important requirement is to minimize the computational complexity of cybersecurity solutions. Considered as a biologically-inspired approach, swarm intelligence optimization is the most promising approach that reduces computational complexity because biological metaphors are an essential part of cyber concepts such as viruses, worms, etc. Bio-inspired approaches can detect a cyberattack in the WSN with very low complexity due to the probabilistic (none-exhaustive) tracking process that intelligently explores the network, and due to the best configuration (i.e. best values of neural network parameters) suggested by the swarm intelligence algorithm, which helps to quickly and efficiently detect any cyberattack.

Conclusion

Cyberattacks continue to become more sophisticated and occur with greater frequency. In this work, we focused on WSN cybersecurity, which is an integral part of many CPSs. In reviewing various bio-inspired approaches to enhance the cybersecurity of CPSs, we found that there is a need to address several of the drawbacks of recently proposed bio-inspired methods. These methods suffer from high computational complexity and require users to choose various input parameters. To address these drawbacks, we proposed SIWC, a generic bio-inspired model that uses a machine learning-based approach. SIWC is an NN system trained by swarm intelligence optimization to automatically determine the optimal critical parameters used to detect cyberattacks.

These methods suffer from high computational complexity and require users to choose various input parameters. To address these drawbacks, we proposed SIWC, a generic bio-inspired model that uses a machine learning-based approach. SIWC is an NN system trained by swarm intelligence optimization to automatically determine the optimal critical parameters used to detect cyberattacks.

Acknowledgments

We thank the anonymous reviewers for their valuable suggestions and comments, which helped improve the quality and presentation of this article.

References

[1] C. Biener, M. Eling, and J. H. Wirfs, "Insurability of Cyber Risk: An Empirical Analysis," *The Geneva Papers on Risk and Insurance–Issues and Practice*, vol. 40, no. 1, 2015, pp. 131–58.

[2] D. S. Ghataoura, J. E. Mitchell, and G. E. Matich, "Networking and Application Interface Technology for Wireless Sensor Network Surveillance and Monitoring," *IEEE Commun. Mag.*, vol. 49, no. 10, 2011, pp. 90–97.

[3] A. Oracevic *et al.*, "Secure Target Detection and Tracking in Mission Critical Wireless Sensor Networks," *Proc. IEEE Int'l. Conf. in Anti-Counterfeiting, Security, and Identification*, 2014, pp. 1–5.

[4] A. Darwish and A. E. Hassanien, "Wearable and Implantable Wireless Sensor Network Solutions for Healthcare Monitoring," *Sensors*, vol. 11, no. 6, 2011, pp. 5561–95.

[5] D. J. John *et al.*, "Evolutionary Based Moving Target Cyber Defense," *Proc. ACM Conf. Genetic and Evolutionary Computation Conf. (GECCO), Wksp. Genetic and Evolutionary Computation in Defense, Security and risk management (SecDef)*, 2014, pp. 1261–68.

[6] S. R. Thompson *et al.*, "Bio-Inspired Cyber Security for Smart Grid Deployments," *Proc. IEEE Innovative Smart Grid Technologies*, 2013, pp. 1–6.

[7] E. W. Fulp *et al.*, "An Evolutionary Strategy for Resilient Cyber Defense," *Proc. IEEE Globecom*, 2015, pp. 1–6.

[8] W. Hao and Z. Yuqing, "AraTRM: Attack Resistible Ant-based Trust and Reputation Model," *Proc. IEEE Int'l. Conf. Computer and Information Technology*, 2014, pp. 652–57.

[9] E. B. Noeparast and T. Banirostam, "A Cognitive Model of Immune System for Increasing Security in Distributed Systems," *Proc. IEEE Int'l. Conf. Comput. Modelling and Simulation*, 2012, pp. 181–86.

[10] F. G Mármol and G. M. Pérez, "Providing Trust in Wireless Sensor Networks Using a Bio-Inspired Technique," *Telecommunication Systems*, vol. 46, no. 2, 2011, pp. 163–80.

[11] R. Khanna, H. Liu, and H. H. Chen, "Reduced Complexity Intrusion Detection in Sensor Networks Using Genetic Algorithm," *Proc. IEEE ICC*, 2009, pp. 1–5.

[12] R. Tanuja *et al.*, "Secure Reputation Update for Target Localization in Wireless Sensor Networks," *Wireless Networks and Computational Intelligence*, Springer, 2012, pp. 109–18.

[13] S. Shamshirband *et al.*, "Co-FAIS: Cooperative Fuzzy Artificial Immune System for Detecting Intrusion in Wireless Sensor Networks," *J. Network and Computer Applications*, vol. 42, 2014, pp. 102–17.

[14] R. V. Kulkarni and G. K. Venayagamoorthy, "Neural Network based Secure Media Access Control Protocol for Wireless Sensor Networks," *Proc. IEEE Int'l. Joint Conf. Neural Networks*, 2009, pp. 1680–87.

[15] W. A. H. Ghanem and A. Jantan, "Swarm Intelligence and Neural Network for Data Classification," *Proc. IEEE Int'l. Conf. Control System, Computing and Engineering*, 2014, pp. 196-201.

Biographies

Salim Bitam (salimbitam@gmail.com) is an associate professor in the Computer Science Department at the University of Biskra, Algeria, as well as a senior member of the LESIA Laboratory at the University of Biskra, and an associate member of the LiSSi Laboratory at the University of Paris-Est Créteil VdM (UPEC), France. He received an Engineer degree in computer science from the University of Constantine, Algeria, his Master's and Ph.D. in computer science from the University of Biskra, and a Doctorate of Sciences (Habilitation) diploma from the Higher School of Computer Science – ESI, Algiers, Algeria. His main research interests are vehicular ad hoc networks, cloud computing, and bio-inspired methods for routing and optimization. He has to his credit more than 30 publications in journals, books, and conferences, for which he has received two best paper awards. He has served as an editorial board member and a reviewer of several journals for IEEE, Elsevier, Wiley, and Springer, and on the technical program committees of several international conferences (IEEE GLOBECOM, IEEE ICC, IEEE/RSJ IROS, and others).

Sherali Zeadally (szeadally@uky.edu) is an associate professor in the College of Communication and Information at the University of Kentucky. He received his Bachelor degree in computer science from the University of Cambridge, England, and his doctoral degree in computer science from the University of Buckingham, England. His research interests focus on computer networks, including wired/wireless networks; network/system/cyber-security; mobile computing; energy-efficient networking; multimedia; and performance evaluation of systems and networks. He is a Fellow of the British Computer Society and the Institution of Engineering Technology, England.

Abdelhamid Mellouk (mellouk@u-pec.fr) is a full professor at the University of Paris-Est Créteil VdM (Paris-12 University UPEC), Networks & Telecommunications Department and LiSSi Laboratory, IUT Creteil/Vitry, France. He graduated in computer network engineering from the Computer Science High Engineering School, University Oran-EsSenia, Algeria, and the University of Paris Sud Orsay (Paris-11 University). He received his Ph.D. in computer science from the same university, and a Doctorate of Sciences (Habilitation) diploma from UPEC. He is the founder of the Network Control Research activity with extensive international academic and industrial collaborations. His general area of research focus is on computer networks, including adaptive real-time bio-inspired control mechanisms for high-speed new generation dynamic wired/wireless networking in order to maintain acceptable quality of service/experience for added value services. He has held several national and international offices, including leadership positions in IEEE Communications Society Technical Committees.

Decapitation via Digital Epidemics: A Bio-Inspired Transmissive Attack

Pin-Yu Chen, Ching-Chao Lin, Shin-Ming Cheng, Hsu-Chun Hsiao, and Chun-Ying Huang

Abstract

The evolution of communication technology and the proliferation of electronic devices have rendered adversaries powerful means for targeted attacks via all sorts of accessible resources. In particular, due to the intrinsic interdependence and ubiquitous connectivity of modern communication systems, adversaries can devise malware that propagates through intermediate hosts to approach the target, to which we refer as transmissive attacks. Inspired by biology, the transmission pattern of such an attack in the digital space much resembles the spread of an epidemic in real life. This article describes transmissive attacks, summarizes the utility of epidemic models in communication systems, and draws connections between transmissive attacks and epidemic models. Simulations, experiments, and ongoing research challenges on transmissive attacks are also addressed.

Introduction

In recent years, researchers have successfully borrowed several biological mechanisms from nature for devising efficient protocols and understanding their performance via the associated mathematical models, especially for cyber security in communication systems [1, 2]. Inspired by epidemiology, this article investigates an emerging attack pattern, *transmissive attack*, featuring heterogeneous propagation paths and specific targets. Analogous to the spread of epidemics in nature, malicious codes act as viruses that are capable of infecting hosts (i.e., electronic devices) via various communication resources, and they can be stealthily transported by intermediate hosts to reach the primary hosts (i.e., targets), which is similar to the biological mechanism known as *host specificity*.

Inevitably, the proliferation of electronic devices equipped with communication capabilities and the penetration of the Internet of Things have created ever increasing security threats that we call *digital epidemics*, which may be even more vital than actual transmissive diseases like Dengue fever, ebola, and SARS due to their cyber transmission and dormant operation nature, and their induced loss in properties and privacy. It is worth mentioning that although the fragility of modern communication systems may seem to be shocking news to the world, the severe consequences caused by digital epidemics have been foreseen by researchers [3–5]. In the past two decades various advanced communication technologies, such as cellular systems and wired and wireless networks, and tremendous user activities, such as online social networking and mobile applications, have constituted a heterogeneous but ubiquitous network among users and devices around the globe, which is known to be a complex communication network [6] or a generalized social network [5]. Malicious codes are able to exploit these heterogeneous communication paths and intrinsic system interconnectivity for propagation, and thereby compromise more devices.

By investigating recently discovered attack cases and system vulnerabilities, we present an emerging attack pattern named *transmissive attack*, where an adversary can leverage diverse communication paths and common communication protocols (e.g., the Internet of Things) to indirectly compromise a target (or a set of targets) that cannot be directly accessed by the adversary.

Furthermore, in order to increase the possibility of reaching the target, transmissive attacks may camouflage their activities to elude detection rather than indiscriminately infesting as many hosts as possible. The specificity to targeted attacks and heterogeneity in propagation paths distinguish transmissive attacks from well-known Internet worms such as Code Red, which only exploits a single propagation resource (the Internet) and features indiscriminate attacks.

Figure 1 illustrates several possible means for the adversary to access the target.[1] Consider the target to be a personal computer in an enterprise that is granted access to private employee/customer databases or confidential corporate files, and all external networking connections to the target are prohibited. If the target is connected to other internal machines that connect to the outside world, the adversary can eventually reach the target by successive propagation. Pessimistically, even if all connections from other internal machines to the target are prohibited, the adversary can still manage to approach the target by compromising the authorized user's electronic devices, such as portable storage devices, wearable

This article describes transmissive attacks, summarizes the utility of epidemic models in communication systems, and draws connections between transmissive attacks and epidemic models. Simulations, experiments, and ongoing research challenges on transmissive attacks are also addressed.

Pin-Yu Chen is with the University of Michigan; Ching-Chao Lin and Shin-Ming Cheng are with National Taiwan University of Science and Technology; Hsu-Chun Hsiao is with National Taiwan University and CTTI, Academia Sinica; Chun-Ying Huang is with National Chiao Tung University.

[1] A detailed attack scenario is provided in the supplementary file, http://arxiv.org/abs/1603.00588

0163-6804/16/$25.00 © 2016 IEEE

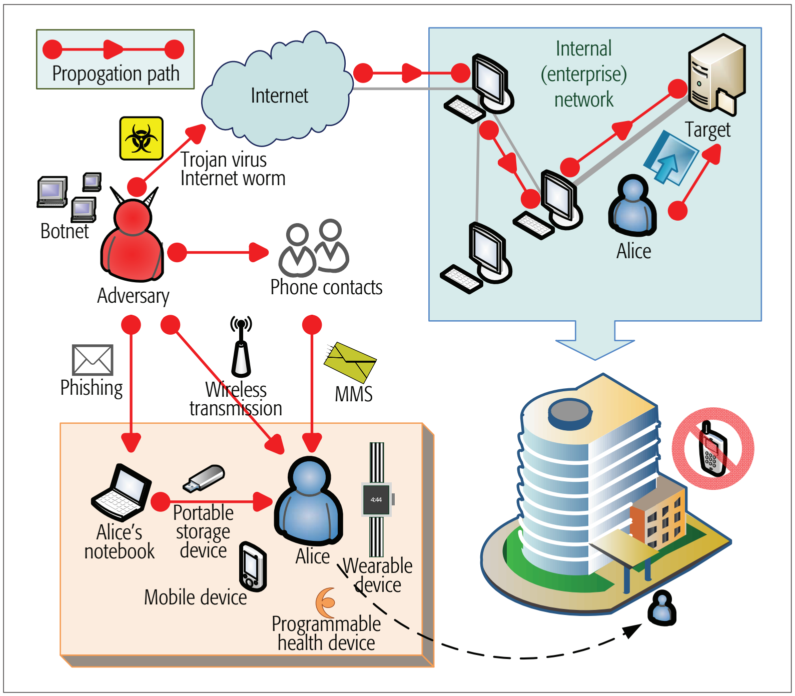


Figure 1. Illustration of transmissive attacks and their propagation paths. Transmissive attacks exploit various communication resources for propagation in order to reach the target. This diagram shows some examples of propagation paths that lead to the target.

devices, or health devices embedded in a human body equipped with communication capabilities. In practice, all the adversary needs to do to launch a transmissive attack is simply release a malicious code (e.g., a Trojan virus), and then sit back and wait for the code to propagate among hosts (potential victims), via either cyber connection (e.g., phishing from the Internet) or human carrier (e.g., Bluetooth or WiFi direct reception from proximity), to create an indirect (i.e., multi-hop) communication path for accessing the target. Moreover, after successful intrusion the adversary can erase its traces from the communication path (e.g., implementing a global timer for self-deactivation) to reduce the risk of being uncovered.

Inspired by biology, we use epidemic models to evaluate the consequences of a transmissive attack from a macroscopic system-level perspective. Analogous to disease transmission assessment, epidemic models categorize the hosts in a system into a few states to analyze the collective behavior of a system with parametric mathematical models (e.g., coupled state difference equations or Markov chains) for the purposes of status tracking, outbreak prediction, and further actions. As a first step toward analyzing transmissive attacks, we use epidemic models to investigate the probability of successfully compromising the target and quantify the risk of exposure with respect to time. We show that the trade-offs in time between the probability of successful intrusion to the target and the associated risk can be characterized by epidemic models, thereby enabling security analysis.[2]

Transmissive Attacks in Practice

As illustrated in Fig. 1, one typical scenario of transmissive attack is that an adversary aims to approach a target by propagating through intermediate hosts via all possible communication resources in a complex communication system. The purpose of such indirect propagation can be that direct access from the adversary to the target is unavailable, or the adversary attempts to hide his/her true identity by manipulating compromised machines to launch an attack, such as the exploitation of mobile devices as botnets [11]. It is worth noting that a transmissive attack can be more insidious due to inherent configurability of electronic devices carried by a user (e.g., programmable in-body health devices or wearable mobile devices), which enables malware propagation even when typical communication devices such as cell phones and laptops are prohibited.

In addition to hidden identity, another appealing advantage of the transmissive attack is that the adversary need not know the complete network topology to accomplish the attack. All the adversary needs to do is release a transmissive malicious code and then wait for the malicious code to propagate to the target due to its transmissive nature. In practice a transmissive attack can be accomplished simply by devising a Trojan virus designed to be operated in the stealthy transmissive mode during propagation and activated when reaching the target. Advanced transmissive attacks can camouflage normal user/network activities to elude intrusion detection or system monitoring, thereby incurring severe threats to security and privacy.

One of the most notable targeted attacks is the Stuxnet attack discovered in 2010. Stuxnet is designed to target a specific version of industrial control systems in a surreptitious manner, whereas traditional worms often aim to infest as many hosts as possible in a short time period. Stuxnet thus exhibits several distinguishing characteristics compared to traditional worms: each Stuxnet worm only replicates itself for at most three times; it is programmed to self-destruct on a day in 2012; it can stealthy propagate via carriers (i.e., vulnerable Windows computers) without showing any symptoms, and only unpack its malicious payload when reaching a target; multiple zero-day vulnerabilities are used. Although performing a targeted attack may be expensive, and indeed Stuxnet is believed to be state-sponsored malware due to its unprecedented level of sophistication, more and more Stuxnet successors (e.g., dudu and Flame) demonstrate how far an adversary is willing to go for high-value targets.

Stuxnet is one kind of advanced persistent threat (APT), which can be seen as one specific case of transmissive attack. An APT possesses the feature of specificity in targets. Although the feature of heterogeneity in community paths is not mandatory for an APT, it would shorten the process to approach the targets if heterogeneous community paths are considered. In 2013, Mandiant[3] summarized the attack life cycle of APT:

1. Initial compromise
2. Establish foothold
3. Escalate privileges
4. Internal reconnaissance
5. Move laterally
6. Maintain presence
7. Complete mission

[2] Although in recent years epidemic-like information propagation has been well studied in the communications society in the context of "epidemic routing," where packets are transmitted in a store-and-forward fashion in intermittently connected networks, little is known on how to apply these well developed analysis tools [7–10] to model transmissive attacks and beyond.

[3] APT1: Exposing One of China's Cyber Espionage Units, http://intelreport.mandiant.com/}

	Number of vulnerabilities		
Application platforms	2013	2014	2015
Adobe Acrobat Reader	66	44	129
Apple iPhone OS	90	120	375
Apple Mac OS X	65	135	384
Apple WatchOS	–	–	53
Google Android	7	11	130
Microsoft Internet Explorer	129	243	231
Microsoft Office	17	10	40
Microsoft Windows 7	100	36	147
Linux Kernel	189	133	77

Table 1. Statistics of vulnerabilities identified on popular applications and platforms.

An APT attack often loops through steps 3 to 6 until it reaches the specific target. These steps are also applicable to transmissive attacks.

To launch a successful transmissive attack, an attacker would also like to increase heterogeneity in community paths (e.g., by exploiting diverse vulnerabilities). The statistics of recently reported vulnerabilities (Table 1) shows the numbers are consistently increasing for most platforms and applications, even for modern mobile and wearable devices. Consequently, various activities and media, including web downloads, document reading, e-mail reading, short messages delivery, Wi-Fi access, Bluetooth access, and NFC contacts, can be used together to deliver malicious payloads and approach targets. By leveraging these existing vulnerabilities, an attacker can even launch transmissive attacks in the background and be invisible to a user.

For example, in July 2015, an unprecedented vulnerability in the Android system called Stagefright was revealed by the cyber security firm Zimperium.[4] Stagefright leverages the vulnerability of the media library to access users' Android devices through a simple multimedia message service (MMS) without users' awareness.[5,6,7] As approximately 80 percent of mobile devices use Android systems, nearly 1 billion devices are potential victims.[3,4,5] By viewing mobile users using different operating systems as hosts with different levels of immunity to a virus, the Stagefright vulnerability behaves like host specificity in epidemiology, as it can compromise users using Android systems.

Overview of Epidemic Models

Here we provide an overview of classical epidemic models that have been applied to communication systems, particularly for modeling information dissemination, malware propagation, and developing the associated control methods.[8]

Following terminologies from biology and epidemiology, each device in a communication system can be categorized into a few states representing its status. The main utility of such an abstraction is that one can leverage epidemic models to simplify complicated interactions among individuals and extract collective information for large-scale analysis and prediction, for example, tracking pandemic spread patterns and predicting their outbreaks in terms of the infected population. A popular analogy is that each device is either in the *susceptible* (S), *infected* (I), or *recovered* (R) state, known as the SIR model.

For epidemic modeling of normal information dissemination dynamics, including routing in communication networks, rumor, news spread in social networks, and so on, an infected individual means he/she carries a certain message (e.g., a data packet) to be delivered, a susceptible individual means he/she does not carry that message but can potentially be infected, and a recovered individual means he/she is immune to the message and hence ignores the message upon reception; for example, in a cooperative relay-assisted network a device in the recovered state will refuse to receive or forward the packet.

For epidemic modeling of malicious codes propagation dynamics, such as privilege escalation or system vulnerability leakage, an infected individual means he/she is compromised by a malicious code and is being leveraged as a warm bed for further propagation or attack (e.g., a botnet). A susceptible individual means he/she is not compromised, but is still vulnerable to the malicious code. A recovered individual means he/she is free of the threats incurred from the malicious code (e.g., securing one's devices via frequent security patch updates).

The following paragraphs introduce three basic epidemic models and relevant control techniques.

SI Model

The SI model assumes each individual is either in the susceptible or infected state. It can be used to estimate the reception performance of a broadcasting protocol or the dynamics of a malicious code. In [4], the authors show that information dissemination in a fully mixed network of dynamic topology and opportunistic links, such as a mobile contact-based network that possesses time-varying traces due to mobility and temporal connections due to opportunistic contacts, can be captured by an SI model. In [12], the authors show that the trends of self-propagating Internet worms such as Code Red and Slammer can be successfully predicted by SI models. In [5], the authors use the SI model to formulate malware propagation in a hybrid network composed of a social network and a proximal network, where malware can leverage delocalized links (e.g., through MMS) and localized links (e.g., through Bluetooth) for propagation.

SIS Model

Similar to the SI model, the SIS model also assumes that each individual is in either the susceptible or the infected state. The difference is that an SIS model allows an individual to transition from the infected state to the susceptible state. SIS models can be well mapped to the formulation of a typical two-state Markov chain where the steady-state behavior of the entire system is used for analysis. The utility of SIS models can be found in formulating recurrent network

SIS models can be well mapped to the formulation of a typical two-state Markov chain where the steady-state behavior of the entire system is used for analysis. The utility of SIS models can be found in formulating recurrent network behaviors, such as the trends of receiving spam mails, or information dissemination in an evolving environment with system reconfiguration factors.

[4] https://www.zimperium.com/

[5] http://fortune.com/2015/07/27/stagefright-android-vulnerability-text/

[6] http://www.forbes.com/sites/thomasbrewster/2015/07/27/android-text-attacks/

[7] As quoted from Zimperium chief technology officer Zuk Avraham, "These vulnerabilities are extremely dangerous because they do not require that the victim take any action to be exploited. Unlike spear-phishing, where the victim needs to open a PDF file or a link sent by the attacker, this vulnerability can be triggered while you sleep. Before you wake up, the attacker will remove any signs of the device being compromised and you will continue your day as usual – with a trojaned phone." Source: http://venturebeat.com/2015/07/27/researchers-find-vulnerability-that-affects-95-of-android-devices

[8] Due to reference count limitations only a subset of related works are introduced in this section. Interested readers can refer to [7–10, references therein] for more details.

Although transmissive attacks can be a serious threat to cyber security, they are often accompanied by an additional price compared to traditional attack schemes. Notably, their spreading nature and self-propagating patterns enhance the risk of exposure, and hence the attacks may be more likely to be detected.

behaviors, such as the trends of receiving spam mails, or information dissemination in an evolving environment with system reconfiguration factors. In [13], the authors integrate the SIS model with queueing theory to study malware propagation dynamics in a dynamic network.

SIR Model

The SIR model is a widely used model in energy-constrained systems (e.g., a wireless sensor network) or communication systems with control capabilities over information delivery (e.g., a configurable routing protocol). An infected individual can transition from the infected state to the recovered state when certain events occur; for example, a sensor stops forwarding packets due to battery drain. A susceptible node can transition to the recovered state when certain mechanisms are activated; for example, a computer is no longer vulnerable to a malicious code after installing the corresponding security patch or upgrading its operating system. In [14], the SIR model is used to study the vulnerability of broadcast protocols in wireless sensor networks. In [7, 8], the SIR model is used to analyze the performance of several protocols for epidemic routing.

Control Techniques

One major advantage of using epidemic models for modeling dynamics of information delivery or malware propagation lies in the fact that their analytical expressions closely resemble coupled state equations appearing in control theory, which allows one to quantify a cost function of interest and evaluate the performance of a control strategy. A commonly used cost function rooted in various applications is the accumulated infected population within a time interval. For instance, in store-and-forward routing schemes such as epidemic routing, the accumulated infected population from the time when a source releases a packet to the time when the packet is no longer carried by any individual is considered as a cost function for data transmission. It can be interpreted as the system-wise buffer occupancy for data transmission since all infected devices need to keep the packet in their own buffer until the destination successfully receives the packet.

Notably, although epidemic routing enables communications in intermittently connected networks, its spreading nature inevitably induces additional system burden, especially for buffer occupancy. In [7], the authors propose two strategies for controlling buffer occupancy, which we call the *global timeout scheme* and the *antipacket dissemination scheme*. In the global timeout scheme, each infected individual drops the packet in its buffer when the global timer expires. In the antipacket dissemination scheme, as motivated by vaccination from immunology, upon packet reception the destination releases an antipacket as an indicator of acknowledgment (ACK) and asks every encountered individual to forward the antipacket so that infected nodes can erase the obsolete packet from its buffer, and susceptible nodes can be prevented from receiving the already delivered packet, and hence achieve buffer occupancy reduction.

In [9], the authors consider time-dependent control capability of SIR models in hybrid networks, where the control ability is proportional to the elapsed time; that is, the ability to restrain malware propagation increases with the time spent in reverse-engineering its operations. An optimal control strategy based on dynamic programming is proposed for solving the optimal time to implement the control strategy (analogously releasing the antidotes) in order to balance the trade-offs between effectiveness and consequences.

Connecting the Dots: Evaluating Transmissive Attacks via Epidemic Models

Although transmissive attacks can be a serious threat to cyber security, they are often accompanied by an additional price compared to traditional attack schemes. Notably, their spreading nature and self-propagating patterns enhance the risk of exposure, and hence the attacks may be more likely to be detected. Generally speaking, while an attacker can accelerate the processes of reaching the target by compromising additional hosts, such an increased level of malicious activities becomes easily identified, thereby jeopardizing the purpose of the attack. To this end, there is a trade-off between the probability of a successful attack and the risk of being detected due to excessive exposure.

To quantify this trade-off between attack success and risk of exposure for transmissive attacks, we propose to use epidemic models for analysis. The risk of exposure is the accumulated infected population (i.e., accumulated number of compromised hosts) from time 0 when the adversary launches a transmissive attack to time T when the target is comprised, or the time when the adversary decides to abort the attack, as the longer the duration of a host being compromised renders an attack more prone to detection. The attack success at a time instance t is defined as the probability of successfully accessing the target between time intervals 0 and t.

For further illustration, we consider the scenario where the adversary adopts the global timeout scheme for transmissive attacks as his/her control technique to reduce the risk of exposure. A global timer is set since the attacker launches a transmissive attack, and upon global timer expiration all malware residing in the compromised hosts will erase their traces via complete self-deletion, whether the attack is successful or not, so as to allow the adversary to constrain malware propagation and alleviate exposure.

Under the global timeout scheme, an interesting question that naturally arises is: what is the optimal global timeout value T_G such that the attack success at time T_G is no less than a certain value (e.g., 80 percent) while the risk of exposure can be minimized? Partial answers to this question have been given in the contexts of minimizing the system buffer occupancy while simultaneously guaranteeing end-to-end data delivery reliability between a source-destination pair for epidemic routing [15], where the data delivery reliability and buffer occupancy are proven to be associated with the accumulated infected population under the SIR model [8].

In particular, if the mobility pattern follows a homogeneous mixing mobility assumption, such as the random waypoint model or the random direction model, a closed-form expression of optimal global timeout value is provided in [15] in which, given a data delivery reliability guarantee, the optimal global timeout value that minimizes the system buffer occupancy depends on the initially infected population and the pairwise meeting rate. This suggests that if mobile botnets (i.e., several initially compromised hosts) are utilized to launch a transmissive attack, the global timer should be set smaller than that of a single seed to minimize the risk of exposure. Similarly, the global timer should decrease when the pairwise meeting rate is higher due to more frequent encounters facilitating malware propagation. Moreover, an interesting finding in [15] is that the optimal buffer occupancy grows exponentially with the data delivery reliability. Analogously, for transmissive attacks the exponential growth rate suggests that the risk of exposure can be significantly amplified if an adversary desires higher attack success.

It is also proven in [15] that when adopting the optimal global timer, the per-user buffer occupancy does not depend on the total population for epidemic routing. This suggests that for transmissive attacks, the risk of exposure for a single host can be controlled to a certain extent such that its local risk does not increase with the total host number.

Experiments and Simulations

In this section we conduct several simulations and experiments as a first step toward the analysis of transmissive attacks using epidemic models. In particular, we investigate the trade-offs between the attack success and risk of exposure by simulating global-timeout-value-enabled transmissive attacks in mobile networks with two widely adopted mobility models: the random waypoint (RWP) mobility model and the random direction (RD) mobility model. We also evaluate the effect of propagation path diversity of a mobile social network on transmissive attacks based on mobile and social interaction patterns extracted from real-life datasets.

Simulation of Transmissive Attacks in Mobile Networks

We simulate the traces of a mobile network of N mobile users moving around in a wraparound $L \times L$ square area. Any pair of users can exchange information for communication when they are within distance r of each other. For the RWP mobility model each user selects a destination at random and travels to the destination at a constant speed v. Similarly, for the RD mobility model each user selects a direction at random and travels at a constant speed v. For both models the speed v is randomly and uniformly drawn from the interval $[v_{\min}, v_{\max}]$. Initially (at time 0), one user is compromised to launch a transmissive attack, and the target is selected at random.

Figures 2 and 3 display the attack success and the risk of exposure with respect to the global timeout value T_G, respectively. Given T_G, the attack success is defined as the fraction of simulated transmissive attacks that successfully approach the target prior to time T_G among all trials, and the risk of exposure is defined as the accumulated compromised population divided by the total population N. The SIR epidemic model proposed in [15] is used for performance comparison. The successful rate is the probability of infecting a particular host, and the risk of exposure is evaluated using the accumulated infected population.

It can be observed that the global timeout value T_G indeed governs the performance of both attack success and risk of exposure. The simulation results also validate the trade-offs between these two metrics as the enhancement of attack vulnerability often leads to an increased

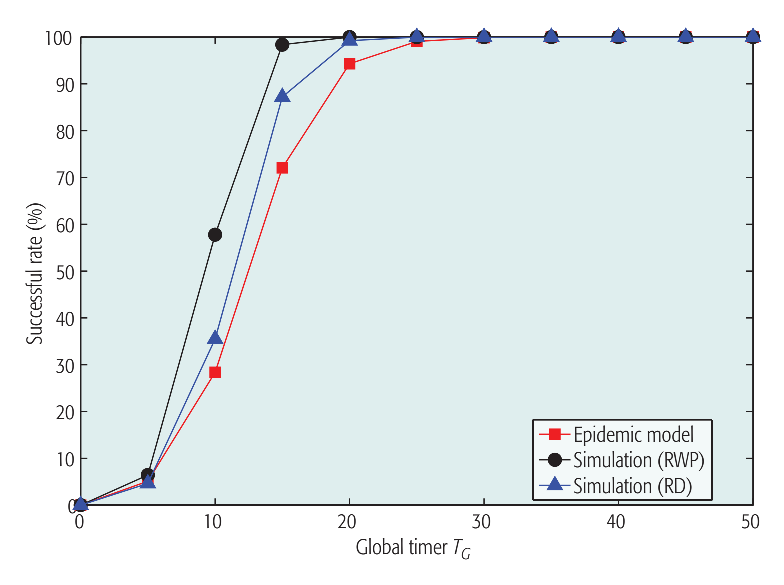


Figure 2. Successful rate for transmissive attacks with respect to varying global timeout value T_G in mobile networks simulated by RD and RWP mobility models. The system parameters are N = 100 mobile users, r = 0.1 km, L = 2.5352 km, $v_{\min}$ = 4 km/h, $v_{\max}$ = 10 km/h, and pairwise meeting rate = 0.37043. The results are averaged over 10,000 trials.

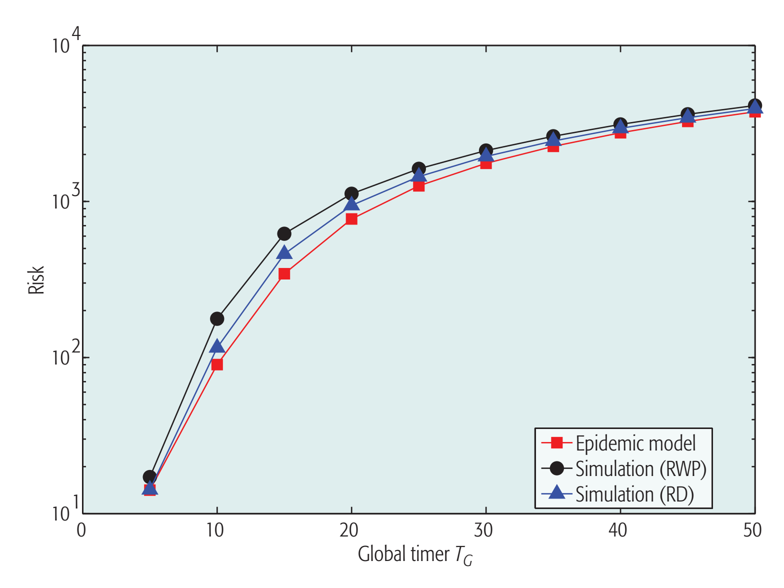


Figure 3. Risk metric corresponding to Fig. 2. Epidemic models are capable of predicting both the attack success and risk of exposure.

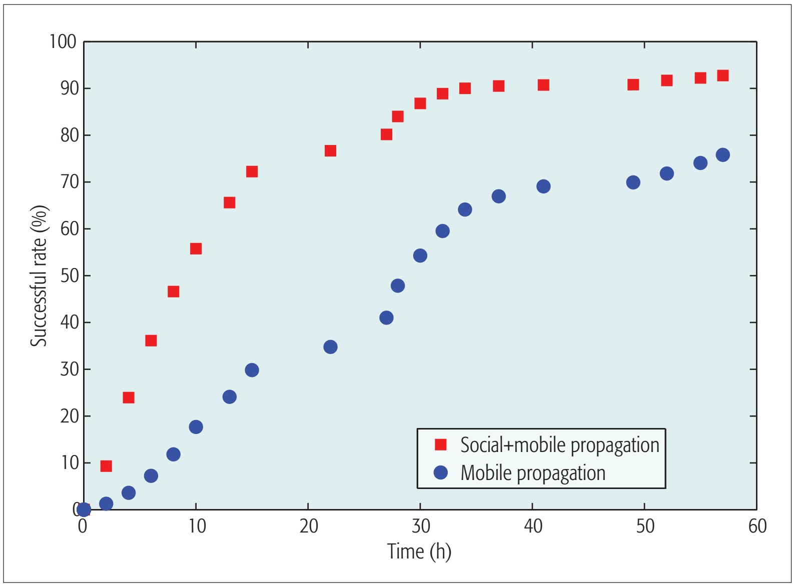


Figure 4. Successful rate for transmissive attacks in a mobile social network: exploiting both social and mobile propagation paths can significantly improve the possibility of approaching the target. The propagation parameters $p_s = p_\ell = 0.05$, and the results are averaged over 10,000 trials.

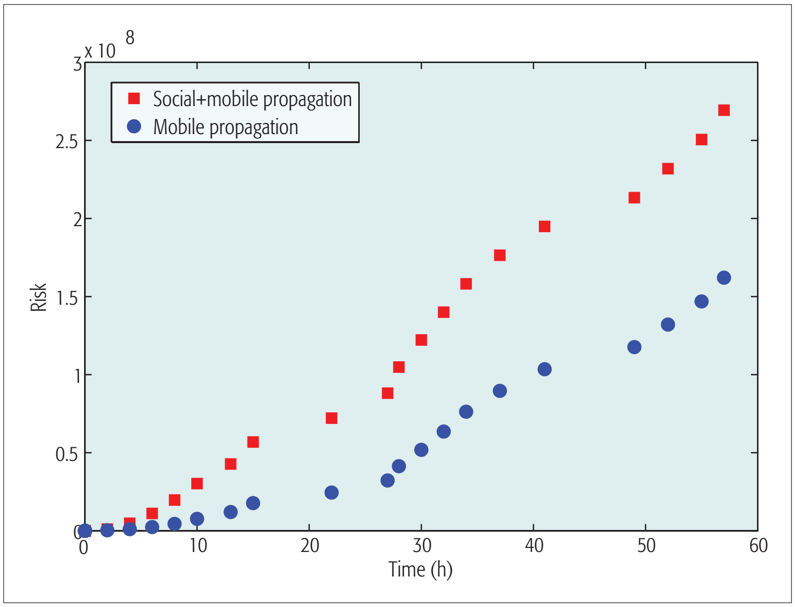


Figure 5. Risk metric corresponding to Fig. 4: the accumulated infected population with respect to time. The results show clear trade-offs between attack success and risk of exposure.

risk of exposure, and vice versa. For example, to enhance the attack success from 30 percent ($T_G = 10$) to 90 percent ($T_G = 20$), the risk of exposure needs to be amplified 10 times. Notably, the predicted results from the epidemic model can successfully capture the trends of these two metrics. An immediate utility is that an adversary can use the epidemic model to determine the optimal global timeout value that guarantees the attack success while simultaneously minimizing the risk of exposure (e.g., selecting $T_G = 25$ such that the attack success is no less than 95 percent). Moreover, a defender can also utilize the epidemic model to evaluate a system's vulnerability without conducting time-consuming simulations.

Experiment of Transmissive Attacks in Mobile Social Networks

To investigate the impact of propagation path diversity on transmissive attacks, we use the CRAWDAD mobile-social interaction traces[9] to simulate a transmissive attack. The purpose of this experiment is to study the consequences of transmissive attacks that are capable of propagating through social contacts (e.g., via MMS) or proximity contacts (e.g., via Bluetooth). In such a mobile social network the malware can propagate from one compromised user to another user with probability of success p_s via the social propagation path if these two users are social contacts (i.e., there is an edge between these two users in the corresponding social graph). Similarly, the malware can propagate from one compromised user to another user with probability of success p_ℓ via the proximity propagation path if these two users are within a physical contact distance.

Figures 4 and 5 display the attack success and risk of exposure for transmissive attacks in the mobile social network, respectively. It can be observed that the inclusion of social propagation paths can significantly enhance the attack success. For example, after 30 hours since launching a transmissive attack, the attack success in utilizing both social and mobile propagation paths can be doubled compared to the attack success of only exploiting mobile propagation paths. However, the induced risk metric is also amplified, as shown in Fig. 5.

Additional experiments of different parameter configurations show similar trends in attack success and risk of exposure, which are discussed in the supplementary file.[1] These results suggest that propagation path diversity can facilitate transmissive attacks at the price of potentially amplified exposure. In addition, how current epidemic models can be improved to model transmissive attacks in such a heterogeneous network is an active research area.

Some Ongoing Challenges and Open Research Questions

Here we discuss several ongoing challenges and open research questions related to transmissive attacks.

Lateral movement detection and prevention: Unlike disruption attacks (e.g., denial of service), which often cause distinguishable anomalous activities, lateral movement attacks (e.g., privilege escalation that insidiously acquires user credentials) are difficult to detect. Transmissive attacks fall into one category of lateral movement attacks due to their stealthy transmissive nature. If detecting lateral movement is implausible, one may shift attention to designing a resilient cyber system that can constrain the damage induced by such attacks.

Transmissive attacks in a network of networks: A network of networks (NoNs) is an intuitive explanation of modern communication systems with intrinsic layered structures and heterogeneous networks. The layers of the Internet architecture can be operated by different protocols, and a device can have multiple communication resources (e.g., cellular, WiFi, and Bluetooth modules).

As horizontal malware propagations within a single layer/system can be straightforward by lever-

[9] CRAWDAD dataset thlab/sigcomm2009 (v. 2012-07-15); http://crawdad.org/thlab/sigcomm2009/20120715

aging similar vulnerabilities, vertical malware propagations traversing different layers/systems can be more difficult due to lack of common vulnerabilities or implementation of additional security rules. In terms of bio-inspired attacks, transmissive attacks that are self-evolving and adaptive to the NoN environment can be a vital threat.

Data-driven inference for attack and defense: In a data-rich era, our cyber footage is everywhere and easy to track. Both attackers and defenders should make use of available data collected from different sources to infer vulnerabilities in a system. Notably, modern technology enables an adversary to optimize his/her attack strategy based on the inference results from collected data prior to launching a transmissive attack, known as the inference attacks. For instance, personal trace information such as GPS signals or locations revealed by online social networking activities can be directly observed or indirectly inferred from user-centric data.

Evolutionary resilience of dynamic systems: In many cases the underlying communication system where a transmissive attack takes place is an ever changing system due to variations in time, traffic flows, evolution of communication technology, and so on. Therefore, a general notion of resilience for such a dynamic system is necessary to quantify network stability that can vary with time, which we call *evolutionary resilience*. Notably, biology models such as ecological systems, predator-prey models, and evolutionary game theory that target evolutionary stability in time-varying coupled systems may be well mapped to analyze transmissive attacks in dynamic systems.

In many cases the underlying communication system where a transmissive attack takes place is an ever-changing system due to variations in time, traffic flows, evolution of communication technology, and so on. Therefore, a general notion of resilience for such a dynamic system is necessary to quantify network stability that can vary with time, which we call evolutionary resilience.

Conclusion

This article introduces an emerging attack pattern called transmissive attack that leverages diverse communication paths to approach a target and accomplish its task. Inspired by biology, we provide an overview of commonly used epidemic models for communication systems, and connect the dots between transmissive attacks and epidemic models. We perform simulations via two widely used mobility models and conduct experiments in mobile social networks to demonstrate the utility of epidemic models for assessing attack success and risk of exposure, and we also discuss some ongoing research challenges and open research questions related to transmissive attacks.

Acknowledgment

This work was supported in part by Taiwan Information Security Center (TWISC), Academia Sinica, and Ministry of Science and Technology, Taiwan, under the grant MOST 103-2221-E-009-230-MY2, 104-2218-E-001-002, 104-2923-E-011-006-MY2, and 104-2221-E-009-200-MY3. The first author would like to thank Mr. Ming-Yann Lee (M.D.) at the Department of Medical Imagine, National Taiwan University Hospital, for his help in clarifying terminologies in biology and epidemiology.

References

[1] W. Mazurczyk and E. Rzeszutko, "Security – A Perpetual War: Lessons from Nature," *IEEE IT Professional*, vol. 17, no. 1, Jan. 2015, pp. 16–22.

[2] S.-M. Cheng and P.-Y. Chen, "Ecology-Based DoS Attack in Cognitive Radio Networks," 2016, arXiv:1603.01315.

[3] P. Wang *et al.*, "Understanding the Spreading Patterns of Mobile Phone Viruses," *Science*, vol. 324, May 2009, pp. 1071–75.

[4] P.-Y. Chen and K.-C. Chen, "Information Epidemics in Complex Networks with Opportunistic Links and Dynamic Topology," *Proc. IEEE GLOBECOM*, Dec. 2010.

[5] S.-M. Cheng *et al.*, "On Modeling Malware Propagation in Generalized Social Networks," *IEEE Commun. Lett.*, vol. 15, no. 1, Jan. 2011, pp. 25–27.

[6] S.-M. Cheng *et al.*, "Diffusion Models for Information Dissemination Dynamics in Wireless Complex Communication Networks," *J. Complex Systems*, 2013.

[7] Z. J. Haas and T. Small, "A New Networking Model for Biological Applications of Ad Hoc Sensor Networks," *IEEE/ACM Trans. Net.*, vol. 14, no. 1, Feb. 2006, pp. 27–40.

[8] X. Zhang *et al.*, "Performance Modeling of Epidemic Routing," *Comp. Net.*, vol. 51, no. 8, July 2007, pp. 2867–91.

[9] P.-Y. Chen, S.-M. Cheng, and K.-C. Chen, "Optimal Control of Epidemic Information Dissemination Over Networks," *IEEE Trans. Cybernetics*, vol. 44, no. 12, Dec. 2014, pp. 2316–28.

[10] S. Peng, S. Yu, and A. Yang, "Smartphone Malware and Its Propagation Modeling: A Survey," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 2, 2014, pp. 925–41.

[11] A. Mtibaa, K. Harras, and H. Alnuweiri, "From Botnets to Mobibots: A Novel Malicious Communication Paradigm for Mobile Botnets," *IEEE Commun. Mag.*, vol. 53, no. 8, Aug. 2015, pp. 61–67.

[12] C. C. Zou *et al.*, "The Monitoring and Early Detection of Internet Worm," *IEEE/ACM Trans. Net.*, vol. 13, no. 5, Oct. 2005, pp. 961–74.

[13] V. Karyotis and S. Papavassiliou, "Macroscopic Malware Propagation Dynamics for Complex Networks with Churn," *IEEE Commun. Lett.*, vol. 19, no. 4, Apr. 2015, pp. 577–80.

[14] P. De, Y. Liu, and S. Das, "An Epidemic Theoretic Framework for Vulnerability Analysis of Broadcast Protocols in Wireless Sensor Networks," *IEEE Trans. Mobile Comp.*, vol. 8, no. 3, Mar. 2009 , pp. 413–25.

[15] P.-Y. Chen, M.-H. Sung, and S.-M. Cheng, "Buffer Occupancy and Delivery Reliability Tradeoffs for Epidemic Routing," 2016, arXiv:1601.06345.

Biographies

Pin-Yu Chen [S'10] (pinyu@umich.edu) received his B.S. degree in electrical engineering and computer science (undergraduate honors program) from National Chiao Tung University, Taiwan, in 2009, and his M.S. degree in communication engineering from National Taiwan University in 2011. He is currently working toward his Ph.D. degree in electrical engineering and computer science at the University of Michigan, Ann Arbor. His research interests include network science, interdisciplinary network analysis, graph clustering and community detection, statistical graph signal processing, cyber security, and their applications to data analysis and communication systems. He is a member of the Tau Beta Pi Honor Society and the Phi Kappa Phi Honor Society, and was the recipient of the Chia-Lun Lo Fellowship. He was also the recipient of the IEEE GLOBECOM 2010 GOLD Best Paper Award, an IEEE ICASSP 2014 NSF travel grant, and an IEEE ICASSP 2015 SPS travel grant.

Ching-Chao Lin (m10415008@mail.ntust.edu.tw) received his B.S. degree in computer science and information engineering from National Taiwan University of Science and Technology, Taipei, in 2015. He is currently working toward his M.S. degree in computer science and information engineering at National Taiwan University of Science and Technology. His research interests include cyber security and wireless networks.

Shin-Ming Cheng [S'05, M'07] (smcheng@mail.ntust.edu.tw) received his B.S. and Ph.D. degrees in computer science and information engineering from National Taiwan University in 2000 and 2007, respectively. He was a postdoctoral research fellow at the Graduate Institute of Communication Engineering, National Taiwan University, from 2007 to 2012. Since 2012, he has been with the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology as an assistant professor. His current research interests include mobile networks, wireless communication, cyber security, and complex networks. He was a recipient of the IEEE PIMRC 2013 Best Paper Award and the 2014 ACM Taipei/Taiwan Chapter K. T. Li Young Researcher Award.

Hsu-Chun Hsiao (hchsiao@csie.ntu.edu.tw) is an assistant professor in the Department of Computer Science and Information Engineering and Graduate Institute of Networking and Multimedia at National Taiwan University. She also holds an adjunct assistant researcher position at Academia Sinica. She completed her B.S. (2006) and M.S. (2008) in electrical engineering at National Taiwan University, and her Ph.D. at Carnegie Mellon University (2014). Her research interests include network security, anonymity and privacy, and applied cryptography.

Chun-Ying Huang [S'03, M'08] (chuang@cs.nctu.edu.tw) is an associate professor in the Department of Computer Science, National Chiao Tung University. He leads the security and systems laboratory at National Chiao Tung University. From 2008 to 2013, he was an assistant professor in the Department of Computer Science and Engineering, National Taiwan Ocean University, and was an associate professor in the same deparment from 2013 to 2016. He received his Ph.D in electrical engineering from National Taiwan University in 2007. His research interests include system security, multimedia networking, and mobile computing. He is a member of ACM and IEEE. He was a recipient of the 2014 ACM Taipei/Taiwan Chapter K. T. Li Young Researcher Award.

Bio-Inspired RF Steganography via Linear Chirp Radar Signals

Zhiping Zhang, Michael J. Nowak, Michael Wicks, and Zhiqiang Wu

The authors provide a tutorial on a novel RF steganography scheme to conceal digital communication in linear chirp radar signals. They provide a review of the linear chirp signal and existing communication systems using chirp waveforms. They discuss how to implement the RF steganography and hide digitally modulated communication information inside a linear chirp radar signal to prevent an enemy from detecting the existence of such hidden information.

Abstract

The chirp signal is one of the first bio-inspired signals commonly used in RF applications where the term chirp is a reference to the chirping sound made by birds. It has since been recognized that birds communicate through such chirping sounds to attract other birds of the same species, to transmit an alarm for specific threats, and so on. However, birds of a different species, or sometime even birds in a different social group within a species, are unable to connect a specific meaning to certain calls — they will simply hear a bird chirping. Inspired by such, this article provides a tutorial on a novel RF steganography scheme to conceal digital communication in linear chirp radar signals. We first provide a review of the linear chirp signal and existing communication systems using chirp waveforms. Next we discuss how to implement the RF steganography and hide digitally modulated communication information inside a linear chirp radar signal to prevent an enemy from detecting the existence of such hidden information. A new modulation called reduced phase shift keying is employed to make the modulated chirp waveform almost identical to the unmodulated chirp signal. Furthermore, variable symbol durations are employed to eliminate cyclostationary features that might otherwise be exploited by an enemy to detect the existence of the hidden information.

Introduction

The chirp signal is among the first bio-inspired signals commonly used in our RF applications: the name chirp is a reference to the chirping sound made by birds. Using chirps, birds conduct sophisticated communication through their bird calls and bird songs. They transmit various messages to engage in courtship, to convey alarm, or even to advertise territories. Interestingly, some bird calls and bird songs can convey messages that are only partially understood by other species, but more detailed information will only be picked up by birds of the same species, sometimes even between only a few individuals [1]. In other words, bird calls and bird songs may contain multiple "layers" of information simultaneously, encoded by employing subtle variations of chirping. Such minor variations can be considered a type of embedded modulation that carries a secure message only understood by its intended receiver. To the unintended receiver, the signal appears to be a regular chirp. Inspired by this, we discuss a novel RF steganography method using linear chirp signals in this article.

In military operations and other covert operations, it is highly desirable to prevent an enemy and other parties from detecting the existence of RF communication. A great deal of research has been done to develop low probability of detection (LPD) RF waveforms by ensuring that the power spectral density is lower than the noise floor through spread spectrum technologies such as direct-sequence spreading spectrum or frequency hopping [2], by exploiting noise-like signals such as chaotic signals to carry information [3], or both [4].

However, there is another interesting and under-investigated approach to camouflage communication. The idea is not to hide the waveform itself through LPD designs, but to hide the communication in another form of RF transmission. We call this RF steganography. In this article, we discuss a novel RF steganography scheme to obscure a communication signal within a linear chirp radar signal. This newly designed chirp signal will serve two purposes simultaneously: it is still an effective radar signal providing range and Doppler measurement to its original radar operators, and it carries secure digital communication information intended for designated receivers. We first provide a tutorial on the linear chirp signal and conventional chirp communication waveforms. Then we discuss how to use new modulation and system designs to make the linear chirp signal with its embedded communication information almost identical to the original radar signal without any indication of the embedded digital modulation. This will enable the radar operator to maintain the same performance such as auto-correlation function and ambiguity function, while also ensuring that the enemy is unable to detect the existence of the hidden communication signal embedded in the radar signal.

The rest of the article is organized as follows. First, we review the linear chirp radar signal and conventional chirp modulation and communication schemes. We then discuss the RF steganography scheme, which uses reduced phase shift keying modulation. Next, we further enhance the security of the scheme by introducing variable

Zhiping Zhang and Zhiqiang Wu are with Wright State University; Michael J. Nowak is with the Air Force Research Laboratory; Michael Wicks is with the University of Dayton.

0163-6804/16/$25.00 © 2016 IEEE

symbol duration and modulation design. The conclusion follows.

Linear Chirp for Radar and Communication

Linear chirp signals are widely used today in sonar and radar systems [5]. It is a signal in which the frequency linearly increases ("up-chirp") or decreases ("down-chirp") with time. In a linear chirp signal, the instantaneous frequency $f(t)$ varies linearly with time: $f(t) = f_0 + kt$, where f_0 is the starting frequency, and k is the rate of frequency increase or chirp rate. When k is larger than 0, the signal is an up-chirp; when k is less than 0, the signal is a down-chirp.

Since the chirp signal spans a very wide bandwidth, it is resistant to narrowband interference. This inherent capability of interference rejection makes the chirp signal very attractive to spread spectrum communication systems as well, where a significant advantage is the low Doppler sensitivity [6]. Thus, it is no surprise that many researchers have conducted research on the use of chirp signals for communication purposes.

The first communication scheme employing chirp signals is called chirp modulation. Chirp modulation was patented by Sidney Darlington in 1954 [7] with significant later work performed by Winkler in 1962 [8]. The idea of chirp modulation is very simple: binary data is transmitted by mapping the bits into up-chirps and down-chirps. For instance, over one bit period 1 is assigned a chirp with positive rate a and 0 to a chirp with negative rate $-a$.

Figure 1 illustrates the concept of chirp modulation by transmitting 3 bits: 1 0 1. Information bit 1 is represented by an up-chirp waveform, and information bit 0 is represented by a down-chirp waveform.

In chirp modulation, a single bit is transmitted on each chirp signal. To increase the data rate and transmit multiple bits on a single chirp, it is natural to use a shorter symbol duration and to modulate multiple information bits sequentially onto the chirp signal [9]. Figure 2 illustrates an example of such a combination of digital binary phase shift keying (BPSK) and chirp modulation. Figure 2a shows the unmodulated chirp signal, Fig. 2b shows the baseband 2-phase amplitude modulated (2PAM) signal where 3 binary bits (1 0 1) are represented by antipodal amplitudes (+1 or −1), and Fig. 2c shows the modulated chirp signal. The modulation can be performed by simply multiplying the baseband 2PAM signal (Fig. 2b) with the original chirp signal. This is equivalent to introducing a 0 or π phase offset to the unmodulated chirp signal at different data symbols. As can be seen in Fig. 2c, phase reversal occurs when two adjacent data bits are different. As a direct result, the BPSK modulated chirp signal is very different from the unmodulated chirp signal. Therefore, the enemy may easily recognize that there is digital modulation in the signal.

To enhance the communication performance, combining pseudo-noise spreading with chirp modulation was proposed. This scheme was first suggested by Baier *et al.* [10] and Kowatsh *et al.* [6, 11] and termed pseudo-noise-chirp (PN-chirp). The idea is also quite straightforward: a pseudo-noise spreading code with N chirps is used to represent one data symbol in a chirp modulated signal. This way, the processing gain N of the spread spectrum is exploited, which offers a significant performance gain to the chirp modulated communication system.

At the receiver side, a matched filter can be utilized to obtain excellent bit error rate (BER) performance for these chirp modulated communication signals. However, these signals were developed purely for communication purposes, not for hiding information. As can be seen from the various versions of modulated chirp signals, it is apparent that digital modulation has taken place and that the signal is significantly different

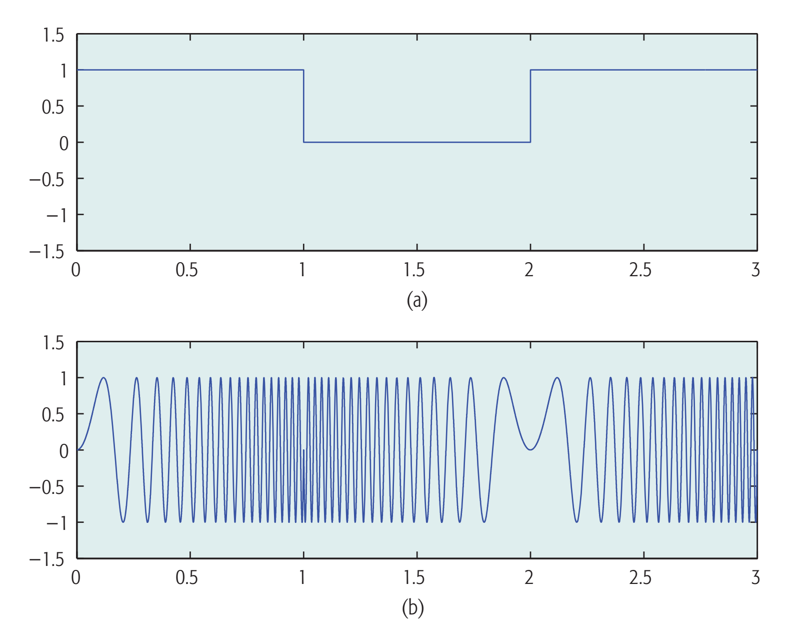


Figure 1. Chirp modulation signal: a) digital data; b) chirp modulation.

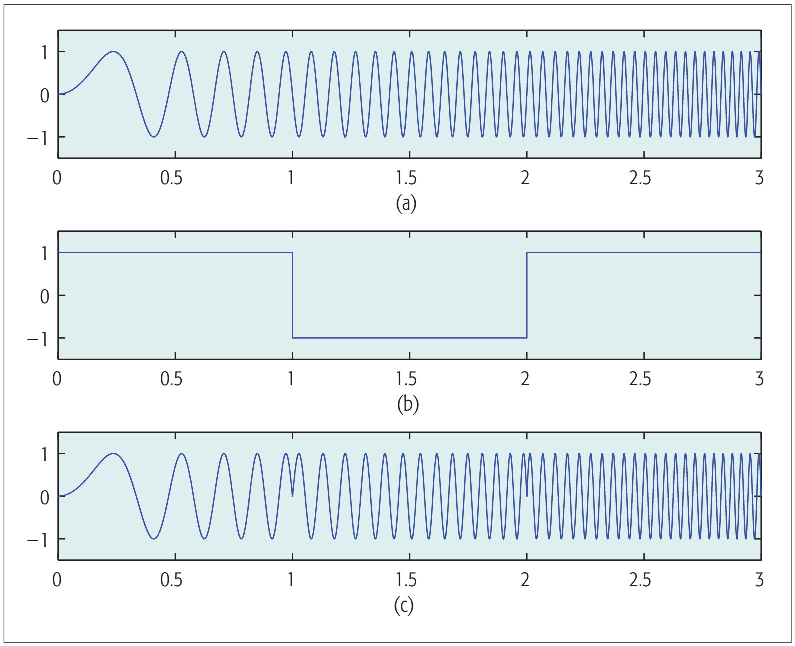


Figure 2. BPSK modulated chirp signal: a) unmodulated chirp signal; b) binary data; c) binary modulated chirp signal.

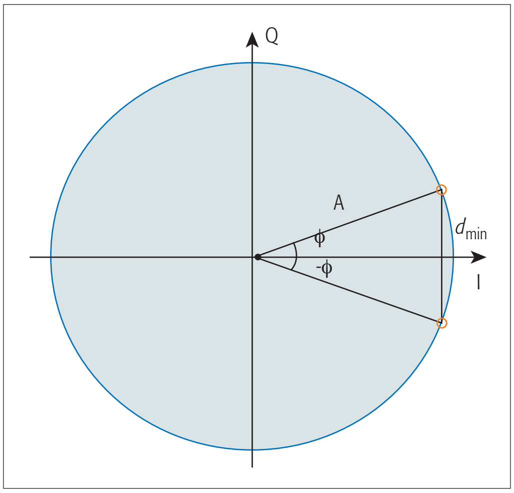


Figure 3. Reduced phase shift keying modulation.

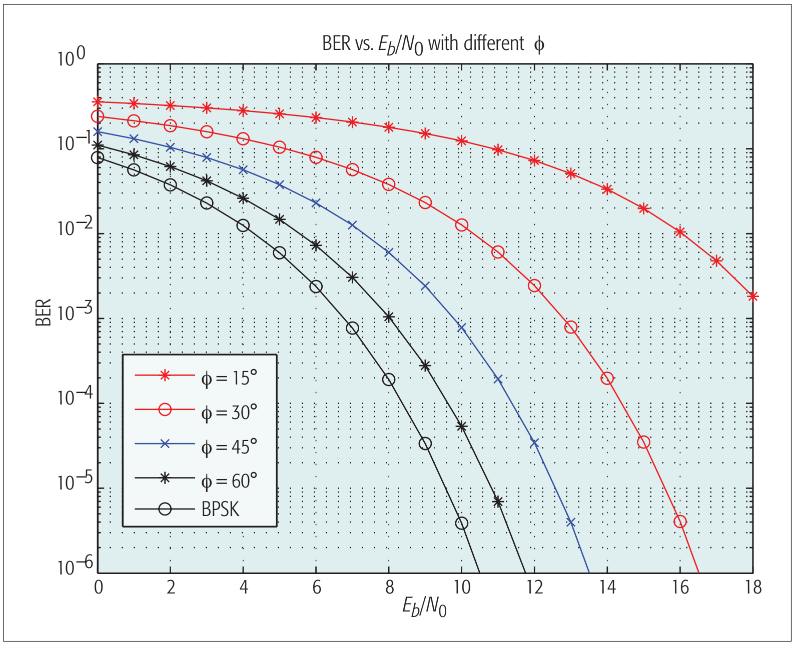


Figure 4. BER of binary reduced phase shift keying.

from an unmodulated chirp signal. Take chirp modulation as an example: a simple spectrogram analysis of the signal will exhibit an up and down frequency change, indicating that the signal is digitally modulated. The BPSK modulated chirp signal and PN-chirp signal exhibit a 180° phase change when two adjacent bits or adjacent chips are different, again revealing the embedded digital modulation.

RF Steganography via Linear Chirp Radar Signal

We now explain RF steganography exploiting a linear chirp radar signal. Pulsed linear chirp signals have been used widely in radar. Specifically, the same linear chirp signal is transmitted from the radar system periodically with a pulse repetition interval larger than the pulse width of the linear chirp signal. In the silent period between two signals, the radar receives the reflections of the previously transmitted signal to perform radar functions such as range detection and Doppler estimation of objects.

Now, instead of designing and transmitting a low probability of detection communication waveform and hoping to avoid detection by the enemy, we embed the communication signal inside the linear chirp radar signal. The joint radar/communication signal is then transmitted from the radar transmitter and reaches the intended communication receiver and also reflects from objects that the radar is designed to detect. The radar signal must complete a round-trip to perform its function, which creates a significant reduction in signal strength. Hence, most radar transmitters emit signals at very high power compared to normal communication transmitters. On the other hand, the embedded communication signal must only make a one-way trip to its target. Therefore, the signal enjoys a very high signal-to-noise ratio (SNR).

This enables us to adopt a different phase shift keying modulation which is suitable for our purposes in this scenario. Instead of using phases 0 and π in the signal constellation to represent our binary data as in regular BPSK, we use two constellation points with a much smaller phase difference (Fig. 3). Specifically, we use phase φ and phase –φ to represent the binary data. We call this new modulation scheme reduced phase shift keying (reduced PSK, or RPSK). As such, the conventional BPSK is a special case of the binary RPSK with φ = 90°.

Because the RPSK modulation uses a smaller phase difference, the constellation points have much smaller separation ($d_{\min}$). As a direct result, this new modulation scheme has poorer BER performance compared to the original BPSK. The BER of the original BPSK in an additive white Gaussian noise (AWGN) channel is determined by

$$Q\left(\frac{d_{min}/2}{\sigma}\right)=Q\left(\frac{A}{\sigma}\right)=Q\left(\sqrt{\frac{2E_b}{N_0}}\right),$$

where $A = E_b$, $N_0/2$ is the power spectral density of the additive white Gaussian noise, and E_b is the bit energy. On the other hand, the BER of binary reduced PSK is

$$Q\left(\frac{d_{\min}/2}{\sigma}\right)=Q\left(\frac{A\sin^2(\phi)}{\sigma}\right)=Q\left(\sqrt{\frac{2E_b\sin^2(\phi)}{N_0}}\right).$$

For example, when φ = 15°, $\sin^2(\phi) = 0.0670$, which corresponds to –11.74 dB. Hence, such a binary RPSK requires 11.74 dB additional SNR to achieve the same BER performance of the original BPSK. However, because of the very high SNR of the linear chirp radar signal, this performance loss is quite tolerable. Figure 4 shows the BER vs. E_b/N_0 curves of binary RPSK modulations with different phase φ. As shown in Fig. 4, there is an 11.74 dB difference between the BER of a binary reduced PSK and that of a conventional BPSK (where φ = 90°).

Additionally, the reduced PSK modulation can be coupled with pseudo-noise code spreading spectrum. A pseudo-noise code with code length N larger than 15 (11.76 dB) will be sufficient to compensate for the performance loss of an RPSK modulation with φ of 15°.

It is important and interesting to note that when we project the binary RPSK constellation to the in-phase component and quadrature component, it is evident that only the quadrature component contains the digital data. Hence, the optimum receiver can be implemented easily by a matched filter with only the quadrature component.

It is also worth mentioning that if a higher data rate is required for the covert communication, we can employ higher M-ary PSK modulations in the RPSK setting by assigning M constellation points with phase difference 2ϕ. For example, a quadrature RPSK (QRPSK) constellation consists of four phases: ϕ, $3\ \phi$, $-\phi$, and $-3\ \phi$.

With the new RPSK modulation, the linear chirp signal with embedded communication much more closely resembles the original unmodulated linear chirp signal. Figure 5 shows the same example of a chirp modulated signal with 3 bits modulated using binary RPSK with $\phi = 15°$. Compared to Fig. 2, it is clear that the phase reversal between adjacent symbols in the original BPSK modulated chirp signal is now replaced by a much smaller phase change, leading to a signal almost identical to the original unmodulated chirp signal. It is evident that the smaller the ϕ is, the less difference there is between the modulated chirp signal and the original chirp signal. When RPSK modulation with small ϕ is applied, neither time domain analysis (e.g., zero crossing rate) nor spectrum analysis (e.g., spectrogram) will indicate that there is a digitally modulated signal hidden in the linear chirp signal.

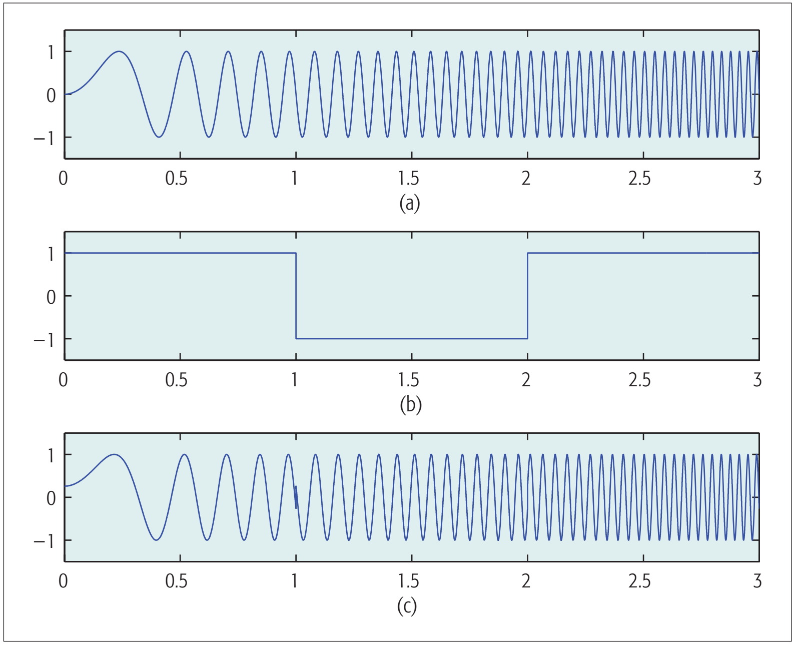


Figure 5. a) Unmodulated chirp signal; b) binary data; c) binary RPSK modulated chirp signal.

Enhanced RF Steganography via Variable Symbol Duration

Despite the improvements already discussed, the chirp signal with RPSK modulation itself is not enough to yield the RF steganography capability we desire. Although the signal does exhibit almost identical time and frequency behavior as an unmodulated linear chirp, it has a cyclostationary feature that is exploitable by an enemy to discover the existence of digital modulation embedded in the radar signal.

Cylostationary analysis has been recognized as an important tool for performing signal detection, RF parameter estimation, and signal classification [12]. A man-made signal such as a communication signal often exhibits cyclostationarity due to the inherent modulation of the communication signal. The second-order spectral moment, also called the spectral correlation function (SCF), will reveal features of the inherent modulation and its parameters in the cyclic frequency domain. For example, the SCF of a BPSK signal will exhibit peaks in the cyclic frequency domain at twice the carrier frequency f_c and at the symbol rate $1/T_b$ [12].

The linear chirp signal with embedded RPSK modulation that we have discussed so far uses a fixed symbol duration T_b to transmit one bit (or one chip). Therefore, cyclostationary analysis of such a signal will exhibit a peak at the symbol rate $1/T_b$ in the cyclic frequency domain (and at multiples of symbol rate m/T_b where m is an integer). Hence, although time domain analysis and frequency domain analysis do not give our enemy any indication of the existence of embedded digital modulation, a cyclostationary analysis will reveal its existence.

To address this, we now propose to use a variable symbol duration scheme for our modulated linear chip transmission. Specifically, we intentionally assign a unique symbol duration T_{b_i} for the ith data symbol. Each symbol has a different duration, and one symbol's duration is not a multiple of another symbol's. This way, we no longer have a fixed symbol rate, and we eliminate the cyclostationarity associated with the symbol rate. Therefore, there will be no cyclostationary feature exploitable by the enemy to discover the existence of our hidden digital modulation. At our intended receiver, on the other hand, because the unique symbol durations are known, there is no difficulty demodulating the data.

However, the variable symbol duration leads to variable symbol energy (E_b) for different data symbols. Since the linear chirp signal is a constant envelope signal, the symbol energy for the ith symbol is simply $A_c^2/2 \cdot T_{b_i}$ where A_c is the amplitude of the chirp signal. Therefore, the data symbols with longer symbol duration will have higher symbol energy and correspondingly better BER performance. This is obviously not desirable.

Fortunately, we have an elegant and simple solution for this problem. By adjusting the phase difference parameter ϕ in the RPSK modulation, we can compensate for the shorter symbol duration with larger ϕ and ensure that the BER stays the same. For example, if the first data symbol has a symbol duration T_{b_1} and a binary RPSK with ϕ_1, and the second data symbol has a symbol duration T_{b_2} and ϕ_2, as long as we make sure that $T_{b_1} \sin^2(\phi_1) = T_{b_2} \sin^2(\phi_2)$, the BER performance does not change.

Figure 6 illustrates such an RPSK modulated chirp signal with variable symbol duration. As shown in Fig. 6b, three different symbol durations are used to carry three binary bits. The first symbol duration $T_{b_1} = 1$, and the associated phase difference $\phi = 15°$. The second sym-

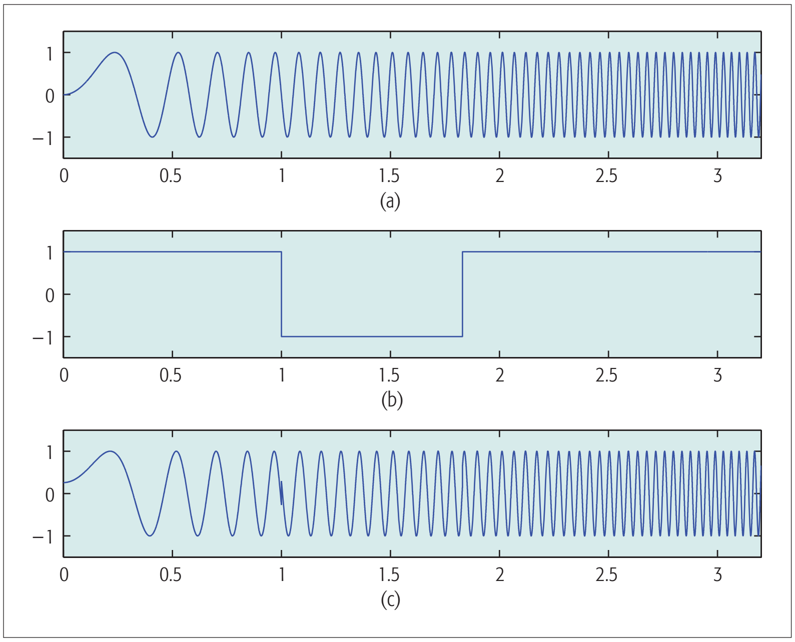


Figure 6. Binary reduced PSK modulated chirp signal with variable symbol duration: a) unmodulated chirp signal; b) binary data; c) binary RPSK modulated chirp signal.

bol duration T_{b_2} = 0.83, and the associated ϕ = 16.5°. The third symbol duration T_{b_3} = 1.37, and ϕ_3 = 12.8°. It is evident that the longer the symbol duration, the smaller the associated ϕ.

Conclusion

In this article, we provide a tutorial on RF steganography by concealing communication information in a linear chirp radar signal. By exploiting a new reduced phase shift keying modulation, the modulated chirp signal becomes very similar to an unmodulated chirp signal. Moreover, by adopting variable symbol durations, we further enhance the security of the hidden communication signal by eliminating the cyclostationary feature. The resulting reduced PSK modulated chirp signal then serves a hybrid radar/communication purpose, while the enemy cannot detect the existence of the modulation embedded in the chirp signal.

Acknowledgment

This material is based on work supported by the National Science Foundation under Grant No. 1323240 and the Air Force Research Laboratory. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] D. E. Kroodsma, *Acoustic Communication in Birds: Volume 1: Production, Perception, and Design Features of Sounds*, Nature Academic Press, 1982.

[2] R. Schoolcraft, “Low Probability of Detection Communications-LPD Waveform Design and Detection Techniques,” *IEEE MILCOM*, 1991.

[3] F. C. Lau and C. K. Tse, *Chaos-Based Digital Communication Systems: Operating Principles, Analysis Methods, and Performance Evaluation*, Springer, 2003

[4] H. Lu *et al.*, “High-Security Chaotic Cognitive Radio System with Subcarrier Shifting,” *IEEE Commun. Letters*, vol. 19, no. 10, Oct. 2015, pp. 1726–29.

[5] N. Levanon and E. Mozeson, *Radar Signals*, Wiley, 2004.

[6] M. Kowatsh and J. T. Laferal, “A Spread-Spectrum Concept Combining Chirp Modulation and Pseudonoise Coding,” *IEEE Trans. Commun.*, vol. com-31, no. 10, Oct. 1983, pp. 1133–42.

[7] S. Darlington, U.S. Patent 2,678,997, Pulse Transmission (chirp).

[8] M. R. Winker, “Chirp Signals for Communications,” *IEEE WESCon Conv. Rec.*, 1962.

[9] C. E. Cook, “Linear FM Signal Formats for Beacon and Communication Systems,” *IEEE Trans. Aerospace Electronic Systems*, vol. AES 10, July 1974, pp. 471–78.

[10] P. W. Baier, R. Simons, and H. Waibel, “Chirp-PN-PSK-Signale als Spread-Spectrum-Signalformen geringer Dopplerempfindlichkeit und grober Signalformvielfalt, NTZ Archiv, vol. 3, Feb. 1981, pp. 29–33.

[11] M. Kowatsch, F. J. Seifert, and J. Lafferl, “Comments on Transmission System Using Pseudonoise Modulations of Linear Chirps,” *IEEE Trans. Aerospace Electronic Systems*, vol. AES-17, Mar. 1981, pp. 300–03.

[12] W. A. Gardner, W. A. Brown, and C.-K. Chen, “Spectral Correlation of Modulated Signals: Part II – Digital Modulation,” *IEEE Trans. Commun.*, vol. 35, no. 6, 1987, pp. 595–601.

Biographies

Zhiping Zhang [M'16] received his B.S. degree in electrical engineering from Nankai University, Tianjin, China, in 2001, and his M.S. and Ph.D. degrees in intelligence science from Peking University, Beijing, China, in 2004 and 2011, respectively. From 2011 to 2013, he was a postdoctoral research fellow at the Department of Computer Science and Technology, Peking University. Since 2013, he has served as a research faculty member and co-director of the Broadband, Mobile and Wireless Networking Laboratory at the Department of Electrical Engineering of Wright State University.

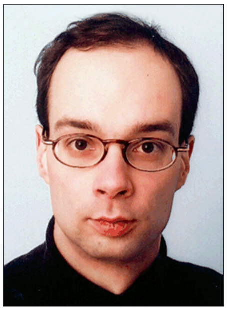
Michael J. Nowak [M'16] received his B.S. in astronautical engineering from the United States Air Force Academy in 1979, his M.S. in aerospace engineering from the University of Texas at Austin in 1992, and his Ph.D. in engineering from Wright State University in 2016. Currently he serves as a technical advisor of the Spectral Warfare Division of Sensors Directorate, Air Force Research Laboratory.

Michael Wicks [S'81, M'89, SM'90, F'98]) received his B.Sc. degree in electrical engineering from Rensselaer Polytechnic Institute, Troy, New York, in 1981, and his M.Sc. and Ph.D. degrees in electrical engineering from Syracuse University, New York, in 1985 and 1995, respectively. He was the U.S. Air Force Senior Scientist for Sensors Signal Processing, specializing in the science and technology needed for superior air and space systems for intelligence, surveillance, reconnaissance (ISR), precision engagement, and electronic warfare. He Joined the University of Dayton in 2011 where he serves as a full professor and Endowed Chair, Ohio Scholar for Sensor Exploitation and Fusion, and Distinguished Research Scientist. He was the recipient of the 2013 IEEE Picard Medal for Radar Technologies and Applications

Zhiqiang Wu [M'02] (zhiqiang.wu@wright.edu) received his B.S. degree from Beijing University of Posts and Telecommunications, China, in 1993, his M.S. degree from Peking University, Beijing, in 1996, and his Ph.D. degree from Colorado State University, Fort Collins, in 2002, all in electrical engineering. He worked as an assistant professor in the Department of Electrical Engineer-ing of West Virginia University Institute of Technology from 2003 to 2005. He joined the Department of Electrical Engineering, Wright State University in 2005, where he currently serves as a full professor.

CONSUMER COMMUNICATIONS AND NETWORKING

Ali C. Begen



Mario Kolberg

Madjid Merabti

This theme is strongly reflected in the four articles we have selected for this edition of the Consumer Communications and Networking series. The theme includes the enhancement of network features as well as network extensions to cover an even larger area and to reach more people. Articles in this edition cover network topics including flow updates in software defined networks, interconnecting ISP content delivery networks, providing connectivity to rural communities, and supporting consumer services on industrial deterministic networks.

More specifically, the first article in this edition, by Yujie Liu *et al.*, focuses on the resource trade-off of flow updates in software defined networks. It discusses how dynamic rule updates affect traffic flow and what additional resources are required to handle the flow update. The article provides both qualitative analysis and quantitative simulation results of the trade-off between bandwidth and flow table size.

The second article, by Yonghwan Bang *et al.*, presents an approach to interconnecting Internet service provider (ISP) content delivery networks (CDNs) based on the Internet Engineering Task Force (IETF) CDN interconnection model. They enhance the IETF work by introducing a CDN interconnection gateway model that overcomes the issue of platform independence between different ISPs. The article reports on a trial carried out between three ISPs in South Korea as the first successful IETF standard-compliant attempt to interconnect ISP CDNs.

The third article, by Saigopal Thota *et al.*, looks at providing connectivity to rural and remote regions. The authors focus on different aspects providing last mile telecommunication connectivity including interfering factors, technology options, and deployment trends. This article serves as a guide to choose, deploy, and operate suitable telecommunications technology depending on the characteristics and features of the area.

The final article in this issue, by Ted Szymanski, focuses on the convergence of industrial networks that provide a deterministic service and the Internet of Things network that provides a best effort service for consumers. The article explores a future Industrial Internet of Things, which is capable of offering both deterministic and best effort services. The article investigates the positive impact of such a network on the delivery of consumer services, large-scale video distribution, e-commerce, and cloud computing.

In closing, we would like to remind you that January is again IEEE CCNC time. The Consumer Communications and Networking Conference will be running for the 14th time between January 8–11, 2017 in Las Vegas, Nevada. IEEE CCNC 2017 will provide a forum to discuss consumer communications issues mentioned in this edition and many more. See http://www.ieee-ccnc.org for details. As in past years, CCNC will run around the same time as the Consumer Electronics Show (CES), giving you two opportunities to learn more about and see consumer communications in action. We hope to see you in Las Vegas in January!

BIOGRAPHIES

ALI C. BEGEN [SM] (acbegen@mediamelon.com) joined MediaMelon, Inc. as the principal architect for streaming technologies in February 2016, where he is currently heading the development efforts for MediaMelon's content-aware streaming solutions. He holds a Ph.D. degree in electrical and computer engineering from Georgia Tech. In January 2016, he was elected a Distinguished Lecturer by the IEEE Communications Society. Visit http://ali.begen.net for further information on his projects, publications, keynotes, tutorials, and teaching, standards, and professional activities.

MARIO KOLBERG [SM] (mkolberg@ieee.org) is a senior lecturer of Computing Science at the University of Stirling. His research interests include P2P overlay networks and home automation. He is on the Board of the *Springer Journal of Peer-to-Peer Networking and Applications* and has a long standing involvement with the IEEE CCNC conference series. He served as its TPC Chair in 2011. He also chaired the Human Centric Computing track at IEEE GLOBECOM 2014. He holds a Ph.D. from the University of Strathclyde, United Kingdom.

MADJID MERABTI [SM] (mmerabti@sharjah.ac.ae) is a professor of networked systems and Dean of the College of Sciences, University of Sharjah, United Arab Emirates. He holds a Ph.D. from Lancaster University, United Kingdom. He has over 25 years' experience in conducting research in the areas of computer networks, mobile computing, and computer network security. He has over 200 publications in these areas and helped in the inception of the IEEE CCNC conference series. He sits on the Editorial Boards of *Wiley Security and Communications Networks Journal* and *Computer Communications Journal*.

On the Resource Trade-off of Flow Update in Software-Defined Networks

Yujie Liu, Yong Li, Yue Wang, Ying Zhang, and Jian Yuan

The authors analyze different existing mechanisms of flow update from their resource utilization perspectives. They study the impact of bandwidth and flow table size on the performance of flow update, and their interactions. They provide both qualitative analysis of the trade-off between these two kinds of resources, and the quantitative simulation results of this trade-off under different realistic network topologies.

ABSTRACT

In software-defined networks, packet forwarding is performed by installing rules on the switches' flow tables. After the rules are installed, the controller needs to dynamically update the rules during runtime for a variety of reasons including traffic engineering, policy changes, network maintenance, and so on. A single forwarding policy update often consists of rule modifications on multiple switches simultaneously. Since the update process requires moving flows to different paths in a consistent and correct manner, multiple steps are usually involved. Thus, additional resources are needed to handle the flow update to ensure correctness and performance, such as extra bandwidth and flow table entries. In this work, we analyze different existing mechanisms of flow update from their resource utilization perspectives. Specifically, we study the impact of bandwidth and flow table size on the performance of flow update, and their interactions. We provide both qualitative analysis of the trade-off between these two kinds of resources, and the quantitative simulation results of this trade-off under different realistic network topologies. Our observation is important to the problems related to flow update, based on which we further illustrate its usefulness with three applications.

INTRODUCTION

With significant momentum in growth, software-defined networking (SDN) has demonstrated efficient control and simplified network management, thanks to the separation of the control and data planes. The SDN controller can centrally and directly configure the forwarding paths of the data plane [1, 2]. While the rules can be installed by the controller proactively in advance, in a production SDN environment, they are updated dynamically and continuously in order to meet the volatile traffic demand, perform scheduled maintenance, and react to emergency situations. A flow update is defined as installing a set of rules on multiple switches to realize a new network policy and to replace an old policy, which is implemented by the controller to replace old forwarding rules with new ones.

Flow update has been studied in a number of existing works lately [3, 4]. We summarize that a good flow update mechanism should complete the update process *fast* and *consistently* with *low overhead*. First of all, due to the dynamic network environment, flow update needs to be carried out frequently, which typically involves the transition of a large number of flows. Thus, during the update, in order to adapt to new situations quickly and reduce the possible network congestion, it is important to complete the update process as fast as possible. Second, the update needs to be consistent, which means that all packets should be processed under either the old policy or the new policy completely. If the packets are processed by a mixture of the two policies, they may be delivered incorrectly or even dropped. Finally, the update process should introduce low overhead to the network, that is, the utilization of network resources cannot exceed the resource limits. For example, an update may accidentally move a large amount of traffic to a loaded link temporarily, which incurs significant overhead. However, if some of the original flows passing through the link are moved first, the overhead may be reduced.

In practice, how to schedule the update of multiple flows to satisfy the above properties is a challenging problem. To complete the flow update fast, many solutions require configuring multiple switches simultaneously. To guarantee consistency, one would need to move the flows from the current state to temporary intermediate states. To reduce overhead, we need to optimize the flows' update sequence to utilize the network resources efficiently. In particular, there are mainly two types of resources. To handle the flows in the intermediate states, network bandwidth is a major constraint [5]. In order to keep the network congestion-free, the link load should not exceed its capacity limit. We assume that the flows' traffic does not change during the update process. However, the update operations could introduce bandwidth overhead. For example, if we migrate a number of new flows into a link before moving its old flows away, the link utilization could get significantly higher than that in the initial and final states. Thus, without careful coordination, the update will result in unfair network bandwidth usage, and even lead to severe traffic congestion and packet loss. On the other hand, from the perspective of forwarding nodes, flow table capacity is another important constraint [6]. It is very limited, since the commonly used flow table, especially ternary content addressable memory (TCAM), is expensive and power hungry. The analysis in [7] demonstrates that in order to use 15-shortest path, up to 20,000 flow entries are required, which is beyond the flow table capacity of even next generation SDN switches. However, to implement a consistent update, the

Yujie Liu, Yong Li, Yue Wang, and Jian Yuan are with Tsinghua University; Ying Zhang is with Ericsson Research Silicon Valley Lab.

0163-6804/16/$25.00 © 2016 IEEE

switches have to carry both the new and old rules in the intermediate states [6]. Thus, in the worst case the flow table space is doubled, which may exceed the capacity of some key switches belonging to multiple routing paths. Once the flow table is fully occupied, the switch may refuse to install other flow entries or even drop the rules silently, which leads to network forwarding errors.

Furthermore, these two constraints of link bandwidth and flow table are closely coupled and should be jointly considered, since rescheduling the traffic load during the flow update has a direct impact on both link utilization and flow table usage concurrently. For instance, transmitting the packets through a path requires sparing the bandwidth of the links as well as adding flow entries into the switches. Besides, splitting the traffic into multiple paths has been used to avoid congestion in the literature [8, 13], but at the cost of more flow entries required in the switches. Therefore, revealing the interactions between these two kinds of resources could provide important hints on the characteristics of network performance during the update, and open up a new avenue for the design of flow update schemes.

In this work, we aim to investigate how the resources of link bandwidth and flow table memory influence the update process, and further analyze how they interact with each other. Through qualitative analysis, we provide a profound understanding of the trade-off between these two resources. By quantitative assessment, we conduct simulations under real network settings to numerically depict the resource trade-off in the flow update. Based on these observations, we further discuss how to solve some related problems in SDN from this new perspective, including update feasibility analysis, heuristic algorithm design and network provisioning in the cloud. Thus, our study provides new insight and ideas for future work in this area.

Flow Update and Existing Solutions

As described above, in the flow update process we should ensure that the link bandwidth and flow table memory limits are not violated. If directly moving a large flow to its new path leads to severe congestion on the bottleneck link, a viable solution is to split the flow into multiple subflows and move the subflows step by step to complete the update [6, 7]. In this way, the link utilization is reduced and the update is carried out successfully, but at the cost of extra flow entries added in the switches. Conversely, if flow table capacity is strictly limited and there is not enough space to configure the rules for flow splitting, we should move the entire flow to reduce the number of extra flow entries. As a result, the traffic load will be unbalanced, or some links will even be overloaded. Thus, link bandwidth and flow table memory are two closely correlated resources, and there is a trade-off between them in the flow update process.

To the best of our knowledge, this trade-off has not been investigated before. However, some recent works [5–10] have studied a part of this problem. We divide these works into three groups, including those focusing on reducing flow table overhead, ensuring a congestion-free update, and jointly considering these two constraints. Based on the classification, we further discuss how the flow update is performed in these works to meet the constraints.

Work	Congestion-freedom	Flow table overhead	Generality	Resource tradeoff
ESPRES [9]	×	✓	×	×
Incremental [6]	×	✓	✓	×
VM migration [10]	✓	×	×	×
zUpdate [5]	✓	✓	×	×
SWAN [7]	✓	✓	×	✓
Dynamic [11]	✓	✓	✓	×
Optimal [8]	✓	✓	✓	×

Table 1. Comparison of the works focusing on network update schemes in SDN.

We first present a survey of existing flow update mechanisms in Table 1. We compare existing works according to four perspectives:

- Congestion freedom: without violating the link bandwidth requirement at any time
- Flow table overhead: without violating the maximum size of flow tables
- Generality: working with general network topologies and traffic demand
- Resource trade-off: considering the trade-off between bandwidth and table size

We discuss the details of each work below. Perešíni *et al.* [9] designed ESPRES, a runtime mechanism that arranges updates to fully utilize the flow tables of switches without overloading them. They propose to reduce rule overhead in the switches by preferring update operations that remove rules first. Katta *et al.* [6] present an algorithm for incremental consistent network update that divides a global policy into a set of slices and updates one slice at a time. They reduce the rule space overhead by increasing the number of slices. However, since the variable link utilization has an ignorable effect on the update process, it is essential to take the link constraint into account, which is not included in these works.

Some other studies have focused on ensuring a congestion-free update. Ghorbani *et al.* [10] propose a heuristic approach for the flow update planning problem in virtual machine (VM) migration. They study how to guarantee that no link capacity is violated at any time during the update process. Liu *et al.* [5] introduce zUpdate to perform congestion-free network-wide traffic migration during data center network updates. They notice that switch table size limit restricts the network update and propose to handle the critical flows specially to cut down the number of added flow entries. However, the interaction of these two constraints is not analyzed, and zUpdate works only for hierarchical network topologies such as FatTree [12]. Hong *et al.* [7] present a novel technique that leverages a small amount of scratch link capacity to apply congestion-free updates. They compute a multi-step transition plan using a linear programming (LP)-based algorithm, and further post-process the output of the LP to fit into the number of rules available. This work focuses on achieving high link

Flow updates face a trade-off between link and flow table resources. Specifically, if the link bandwidth is limited, more flow table resources are needed to achieve a congestion-free update. On the contrary, if the flow table memory is insufficient, more link bandwidth is required to carry out the updates.

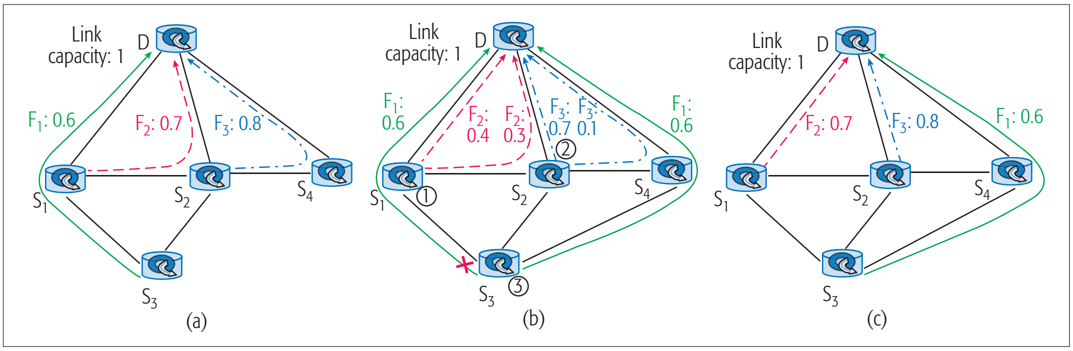


Figure 1. The illustration of the interaction between link capacity and flow table overhead during flow update when link capacity $c = 1$: a) the initial network state; b) the intermediate state; c) the final state. If $c = 2$, the network state can directly transform from a) to c) without splitting traffic, and the flow table overhead is reduced by half compared to b).

utilization in inter-data-center networks, and the intermediate configurations are neither the new nor the old policy in their scenario. Jin *et al.* [11] propose a system for fast and consistent network updates that adapts to runtime conditions. They prove that in the presence of both link capacity and switch memory constraints, finding a feasible update schedule is NP-complete.[1] These works all deal with these two resource constraints separately and do not reveal the interaction between them. Thus, it is important to investigate how to implement flow update by considering this interaction, which will help improve the performance and efficiency of the update algorithm.

The correlation between link bandwidth and flow table capacity has been noticed by several works. Hong *et al.* [7] discuss the fact that switch hardware supports a limited number of forwarding rules, which makes it hard to fully use network capacity. Jain *et al.* [13] find that adding more paths and using fine-grained traffic splitting both give more flexibility to traffic engineering but consume additional hardware table resources. However, this work focuses on centralized traffic engineering, which allocates bandwidth among competing services. In addition, neither of them give a fundamental analysis on the trade-off between these two resources. Liu *et al.* [8] investigate how to carry out a fast update of multiple flows, while this article investigates a different problem of revealing the trade-off between link bandwidth and flow table resource during the update. In [8], both link bandwidth and flow table size constraints are considered in formulating the multi-flow update problem, but the trade-off between these two resources is not taken into account in their proposed flow update algorithms. It analyzes how the flow table size impacts the number of update steps and update success rate, but it does not provide analysis about how many flow entries should be added given the available link bandwidth. Therefore, it is essential to enhance understanding of this resource trade-off by quantitative assessment.

Qualitative Insight

To better explain the resource trade-off in the flow update, we first illustrate it by an example shown in Fig. 1. The packets of three flows, F_1, F_2, and F_3, are sent from S_3, S_1, and S_2 to the same destination D, and their traffic rates are 0.6, 0.7, and 0.8 units, respectively. The initial network state is presented in Fig. 1a. When a new physical link is up between S_3 and S_4, the controller needs to change the forwarding rules to improve routing efficiency, and the new paths are shown in Fig. 1c. If the link capacity is 1 unit, the traffic has to be split to ensure the link utilization is not beyond the capacity limit during the update, which is shown in Fig. 1b. Note that since the number of flow entries installed on each switch increases with the number of flows transmitted through it, large flow table overhead is incurred in the intermediate state. For example, S_4 has to store twice as many flow entries than in the initial state. However, if the link capacity is 2 units, two strategies can be used to implement the update. One is to migrate the entire flow in only one step, which reduces the flow table overhead by half but leads to significantly higher link utilization (75 percent on the link between S_2 and D) and therefore may cause transient congestion. The other is to split the traffic and separate the update into two steps (similar to Fig. 1b), which results in doubling of flow table space compared to the first solution, but reduces the max link utilization to 50 percent.

In conclusion, flow updates face a trade-off between link and flow table resources. Specifically, if the link bandwidth is limited, more flow table resources are needed to achieve a congestion-free update. On the contrary, if the flow table memory is insufficient, more link bandwidth is required to carry out the updates.

Quantitative Evaluation

To quantitatively evaluate the trade-off between the resource utilization of link bandwidth and flow table memory in the update process, we investigate their relationship by employing the solution presented in our earlier work [8] as a case study. The *optimal* solution solves the flow update problem with a multistep strategy. Given the initial and final network configuration, it computes a series of intermediate states to complete the transition by deciding in which step the path of a flow should be updated. Within each step, the two-phase update method [3] is utilized to update the flow table in order to keep per-packet consistency. This mechanism seeks the optimal solution for the flow update problem and aims to minimize the number of update steps, which jointly considers the constraints of flow table and link bandwidth. In the solution, traffic splitting is enabled when moving an entire flow leads to congestion. The mathematical formulation and specific procedures of the algorithm can be found in [8].

[1] NP is the set of problems with solutions that can be verified in polynomial time. A problem p is NP-complete if p is in NP, and every problem in NP is reducible to p in polynomial time. Polynomial time algorithms do not exist for NP-complete problems, unless P = NP, which is still an unsolved question in computer science [14].

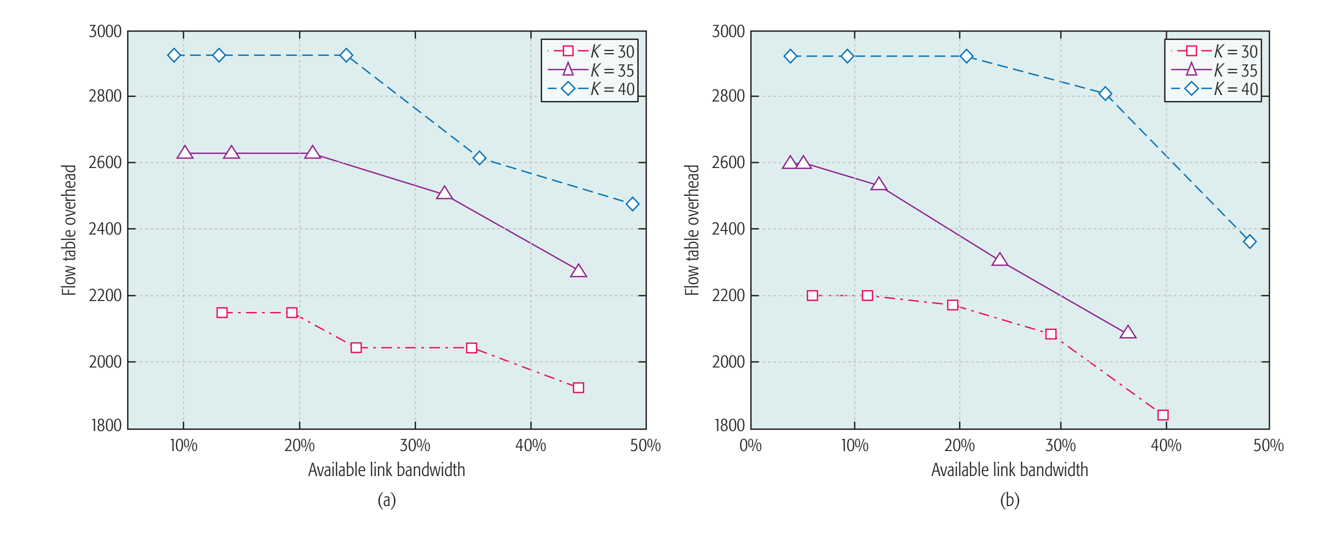


Figure 2. The trade-off between flow table overhead and the available link bandwidth during the updates under different network topology of: a) B4; b) Abilene.

We conduct simulations under two topologies: B4 [13] and Abilene network topology. B4 is Google's inter-data-center WAN, and Abilene is the educational backbone network that connects most universities in the United States. We consider the scenario where the source and destination of the flows are selected randomly, and the initial forwarding paths are calculated using the shortest path in Open Shortest Path First (OSPF) based on the weights assigned to each link. Then we change the weights to simulate events causing flow updates, and obtain the final paths by performing the shortest path method again. The link bandwidth is 1000 Mb/s, and the flow table capacity of all switches is set to be 1000 rules. In order to evaluate how the number of involved flows would impact the trade-off relationship between link bandwidth and flow table overhead during the flow update, we set the number of flows K to be 30, 35, and 40, respectively. We vary the transmission rate per flow, so that the average percentage of available link bandwidth increases from approximately 10 to 50 percent. We measure the total number of flow entries added on the switches during the update as the flow table overhead. The simulation is run 50 times in every case, and the averaged results are calculated.

Figure 2 presents the flow table overhead obtained with variation of the available bandwidth, when the number of updated flows K is 30, 35, and 40, respectively. From the results, we observe that in general the flow table overhead decreases while the available link resource increases. In Fig. 2a, when the number of flows is relatively smaller, the flow table usage varies within a small range. Specifically, when $K = 30$, we can provide 30 percent more link bandwidth to reduce about 9 percent flow table overhead. This trade-off relationship is more obvious as the number of flows rises. When $K = 40$, nearly 40 percent link bandwidth trades for about 17 percent flow table space. We observe similar results in Fig. 2b. When $K = 30$, by increasing the available link bandwidth from 5 to 40 percent, the number of added flow entries is reduced by 18 percent, which is a little more than that in B4 topology. However, when a relatively large number of flows are involved in the update, the flow table overhead is reduced significantly with the increase in link resources. When $K = 40$, 45 percent link bandwidth trades for about 20.6 percent flow table overhead. In general, a certain amount of link bandwidth and flow table space can be exchanged with each other, and the exact amount is related to the network topology and the number of flows. By numerically depicting the trade-off between link utilization and flow table usage, we hope to spark new interest and developments to better orchestrate the flow update process. Several open problems are presented in the next section.

Applications and Open Problems

The trade-off between link utilization and flow table usage helps in deeply understanding the flow update process. From this new perspective, we are able to come up with fast solutions to the problems concerning these two resources in SDN. Three applications are presented in this section.

Flow Update Feasibility Analysis

Before carrying out flow updates, the controller should design a new data plane configuration to realize the new network forwarding policies. If at least one successful flow update plan exists for updating the network to a new configuration, we define this configuration as a feasible one. However, due to the link bandwidth and flow table space constraints, not all of the configurations are feasible in practice. Thus, it is important to carry out feasibility analysis in the selection of the targeted network configuration.

A straightforward method of judging whether a configuration is feasible or not is testing all the possible update sequences, but it is quite time consuming. With the knowledge of the trade-off between link bandwidth and flow table capacity, we can address this problem in a more efficient way. At first, we can derive the number of flow entries that should be added during the update, given the available link bandwidth and the new configuration. Generally speaking, there are two approaches. The first one is to carry out simulations as shown in the previous section to col-

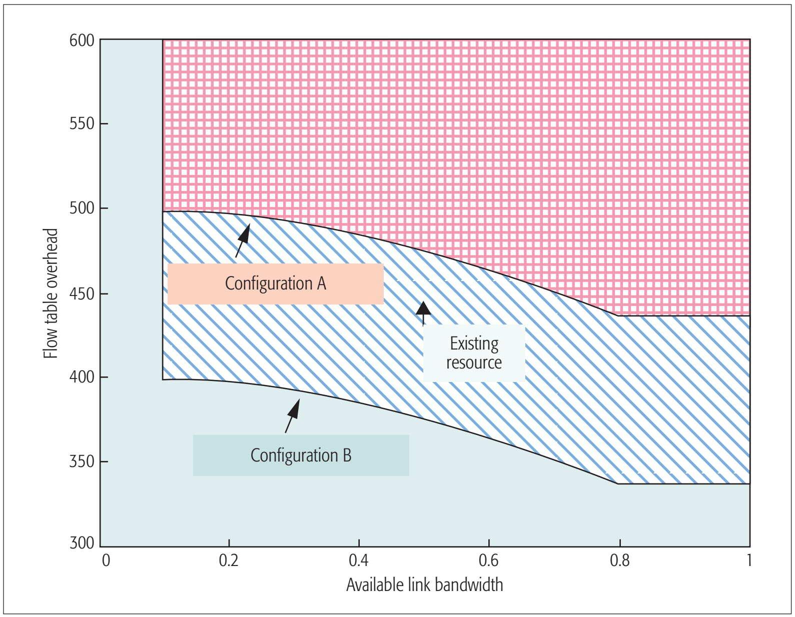


Figure 3. How to utilize the resource trade-off to carry out feasibility analysis of flow updates.

lect data about how many flow entries need to be added based on available link bandwidth. Then these data are fit into curves and the experimental formula is deduced, based on which we can estimate the flow table memory required by the specific flow update scenario. The second one is to build a mathematical model to describe the relationship between the two resources during an update and calculate the number of flow entries to be added based on the formulated model. Since how to build a realistic model and formulate the problem is still an open question that deserves future study, we currently suggest adopting the first approach in practice. If the current flow table memory in the network cannot meet the derived resource requirement, the network configuration is infeasible. As shown in Fig. 3, the top and bottom curves show the resource trade-off when updating the network state to configuration *A* and *B*, respectively. To guarantee that updating to configuration *A* is feasible, the current available network resource should be in the feasible region of *A*, which is represented by the red cross hatched area above the top curve. When the available link bandwidth is the same, updating to configuration *B* incurs less flow table space overhead. Thus, the feasible region of *B* is larger than that of *A*, which includes both the red cross hatched area and the blue backslash hatched area. In this case, assuming the existing resource is denoted by the triangle in Fig. 3, we can conclude that the feasible new configuration is *B* rather than *A*. In conclusion, instead of checking all possible update solutions, the method utilizing the trade-off between link and switch resources improves the efficiency of feasibility analysis in network configuration selection.

Heuristic Algorithm Design

Understanding the relationship between link utilization and flow table usage paves the way for developing efficient flow update algorithms. For example, the problem of finding the fastest update solution of multiple flows without violating the constraints of link bandwidth and flow table is NP-complete [11], because the hardness stems from the fact that flow table constraints involve integers, and flow table space cannot be allocated fractionally. Since polynomial time algorithms do not exist for NP-complete problems (unless P = NP), heuristic algorithms are required for an efficient solution. With the hope of sparking new interest in this area, we discuss how to design an efficient update method of this problem based on the trade-off between the two resources as an example.

At first, we can utilize the derived numerical results regarding the resource trade-off as a benchmark to measure the resource status of the network, which provides the information about which resource is relatively scarce. Then we can decide when and how to schedule the flow update based on the resource status. From the perspective of links, if a link is already in high utilization, an existing flow must be removed before moving a new flow to it. Thus, in deciding the update sequence of the flows, if the link bandwidth is more limited compared to flow table memory, we can give high priority to the update of the critical flows that go through busy links. Otherwise, to ensure that there is enough flow table space to configure the rules for new flows, high priority can be given to updating the flows with paths that include the key switches which have little available flow table memory. When deciding how to update the flows, if the link bandwidth is limited in comparison to flow table space, we may split the flows into multiple subflows and update one subflow to the new path in each step; otherwise, we should update each flow as a whole. In addition, to guarantee consistency, we can adopt existing consistent update mechanisms [3, 4] to update the flow tables. To ensure a fast solution, we update as many flows as possible in one step.

In addition to inspiring heuristic algorithms, understanding the resource trade-off is also important in the formulation of the flow update problem. If we deduce the formula on the numerical relationship of these two resources for the scenario that has typical topologies and specific traffic distribution, the two constraints will be combined. In this situation, the updating solution can be calculated more easily. In conclusion, the idea of considering the resource trade-off opens a new perspective on designing flow update schemes.

Network Provisioning in the Cloud

SDN has attracted increasing attention as a platform for cloud providers. In general, a cloud provider hosts multiple data centers located in different regions. In order to meet users' ever changing requirements, the cloud provider must maintain enough resources to enable necessary flow updates caused by events such as VM migration or traffic engineering. Thus, a fundamental problem for cloud providers is to decide how much link bandwidth and flow table memory need to be provisioned, given the unit price of the resources and the network update requirements.

Now we introduce how to minimize the expenditure of network resources by taking the trade-off between link bandwidth and flow table capacity into account as an example. As shown in Fig. 4, we can derive the numerical relationship of these two resources during the flow update and the feasibility area using the approaches shown in the previous section. If the average flow table memory and link bandwidth of the network is denoted by s and b, respectively, the resource trade-off function is represented by $s = h(b)$, which reveals how much flow table memory should be provided at least under the variation of available link bandwidth. Our goal is to minimize the total cost of network resources, which is represented by $c = f(b, s)$. By reforming this function into $\tilde{s} = g(b, c)$ and combining the trade-off function, the problem is converted to finding the tangent point $P = [b^*, s^*]$ of s and $\tilde{s}$. If $\tilde{s} = g(b, c)$ is linear, such as functions S_1 and S_3 in Fig. 4, P will be one of the two endpoints; otherwise (e.g., S_2), P is somewhere between the endpoints, and should be calculated case by case. In summary, a profound understanding of the trade-off between link and flow table resources will help cloud providers to efficiently decide how many resources should be provisioned in order to reduce the construction cost of data centers.

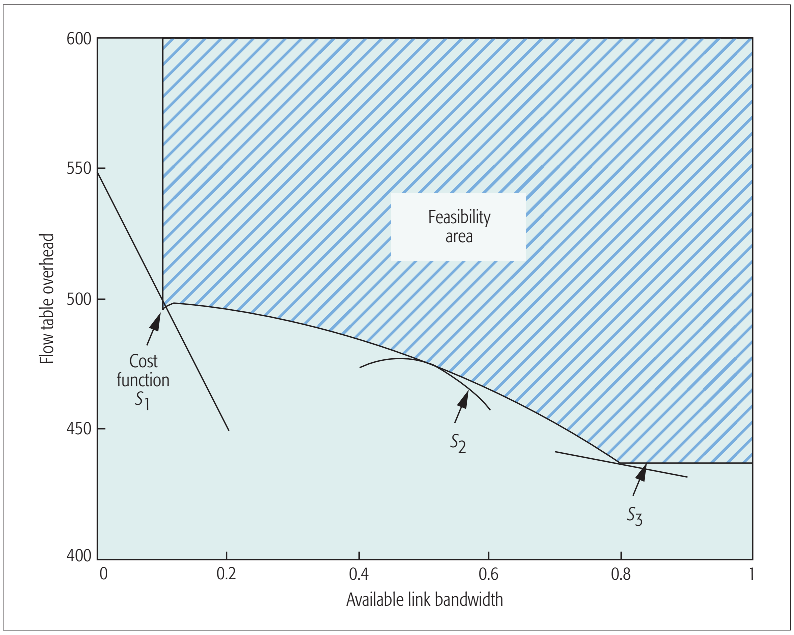


Figure 4. How to utilize the resource trade-off to help cloud providers make informed decisions on resource provisioning.

Conclusion

We have qualitatively analyzed how the resources of link bandwidth and flow table influence the update process, and how they interact with each other. Moreover, we quantitatively evaluate the resource trade-off in the real network update scenarios. Based on these observations, we further discuss how to solve the problems related to flow updates from this new perspective. The analysis results indicate that the efficiency of carrying out feasibility analysis of flow update and performing network provisioning in the cloud can be improved by considering the trade-off, and it also opens up a new avenue for designing heuristic flow update schemes. Thus, our study provides new perspectives and insight for future work in this area.

Acknowledgment

This work is supported by the National Basic Research Program of China (973 Program) under grant 2013CB329105, and the National Nature Science Foundation of China (Grant No. 61273214). This work is also supported by the National Nature Science Foundation of China (No. 61301080) and the Research Fund of Tsinghua University (No. 20161080099).

References

[1] N. McKeown *et al.*, “OpenFlow: Enabling Innovation in Campus Networks,” *SIGCOMM Comp. Commun. Rev.*, vol. 38, no. 2, Apr. 2008, pp. 69–74.

[2] H. Kim and N. Feamster, “Improving Network Management with Software Defined Networking,” *IEEE Commun. Mag.*, vol. 51, no. 2, Feb. 2013, pp. 114–19.

[3] M. Reitblatt *et al.*, “Abstractions for Network Update,” *Proc. ACM SIGCOMM 2012*, Helsinki, Finland, Aug. 13–17, 2012, pp. 323–34.

[4] R. McGeer, “A Safe, Efficient Update Protocol for Openflow Networks,” *Proc. ACM HotSDN 2012*, Helsinki, Finland, Aug. 13–17, 2012, pp. 61–66.

[5] H. H. Liu *et al.*, “zUpdate: Updating Data Center Networks with Zero Loss,” *Proc. ACM SIGCOMM 2013*, Hong Kong, China, Aug. 12–16, 2013, pp. 411–22.

[6] N. P. Katta, J. Rexford, and D. Walker, “Incremental Consistent Updates,” *Proc. ACM HotSDN 2013*, Hong Kong, China, Aug. 16, 2013, pp. 49–54.

[7] C. Y. Hong *et al.*, “Achieving High Utilization with Software-Driven WAN,” *Proc. ACM SIGCOMM 2013*, (Hong Kong, China), Aug. 12–16, 2013, pp. 15–26.

[8] Y. Liu *et al.*, “Achieving Efficient and Fast Update for Multiple Flows in Software-Defined Networks,” *Proc. ACM DCC 2014*, Chicago, IL, Aug. 18, 2014, pp. 77–82.

[9] P. Perešíni *et al.*, “ESPRES: Transparent SDN Update Scheduling,” *Proc. ACM HotSDN 2014*, Chicago, IL, Aug. 22, 2014, pp. 73–78.

[10] S. Ghorbani and M. Caesar, “Walk the Line: Consistent Network Updates with Bandwidth Guarantees,” *Proc. ACM HotSDN 2012*, Helsinki, Finland, Aug. 13–17, 2012, pp. 67–72.

[11] X. Jin *et al.*, “Dynamic Scheduling of Network Updates,” *Proc. ACM SIGCOMM 2014*, Chicago, IL, Aug. 17–21, 2014, pp. 539–50.

[12] C. E. Leiserson, “Fat-trees: Universal Networks for Hardware-Efficient Supercomputing,” *IEEE Trans. Computers*, vol. C-34, no. 10, Oct. 1985, pp. 892–901.

[13] S. Jain *et al.* , “B4: Experience with A Globally-Deployed Software Defined WAN,” *Proc. ACM SIGCOMM 2013*, Hong Kong, China, Aug. 12–16, 2013, pp. 3–14.

[14] C. S. Calude, E. Calude, and M. S. Queen, “Inductive Complexity of the p Versus np Problem,” *Parallel Processing Letters*, vol. 23, no. 1, 2013, pp. 1–16.

Biographies

Yujie Liu received her B. S. degree from the Department of Electronic Engineering, Tsinghua University, China, in 2011. Now she is pursuing her Ph.D. degree at the same university. Her research fields include future Internet architecture, software defined networks, and network functions virtualization.

Yong Li [M’09] received his B.S. and Ph.D degrees from Huazhong University of Science and Technology and Tsinghua University in 2007 and 2012, respectively. During 2012 and 2013 he was a visiting research associate with Telekom Innovation Laboratories and Hong Kong University of Science and Technology, respectively. From 2013 to 2014 he was a visiting scientist at the University of Miami. He is currently a faculty member in the Department of Electronic Engineering, Tsinghua University. His research interests are in the areas of mobile computing and social networks, urban computing and vehicular networks, and network science and future Internet. He has served as General Chair, Technical Program Committee (TPC) chair, and TPC member for several international workshops and conferences. He is currently an Associate Editor of the *Journal of Communications and Networking* and EURASIP *Journal of Wireless Communications and Networking*.

Yue Wang received his B.S. and PhD degrees from the Electronic Engineering Department of Tsinghua University in 1999 and 2005, respectively. He is now an assistant professor at Tsinghua University. His research interests include computer networks, data fusion, and complex networks.

Ying Zhang received her Ph.D. degree from the Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, in 2009. She is a researcher at the IP and Transport Research Group, Ericsson Research Silicon Valley, San Jose, California. Her research interests include networking and systems, including software-defined networking, cloud, Internet, and cellular network management, Internet routing, and measurement and network security.

Jian Yuan received his Ph.D. degree in electrical engineering from the University of Electronic Science and Technology of China in 1998. He is currently a professor in the Department of Electronic Engineering at Tsinghua University. His main research interest is in complex dynamics of networked systems.

CDN Interconnection Service Trial: Implementation and Analysis

Yonghwan Bang, June-Koo Kevin Rhee, KyungSoo Park, Kyongchun Lim, Giyoung Nam, John D. Shinn, Jongmin Lee, Sungmin Jo, Ja-Ryeong Koo, Jonggyu Sung, Young-il Seo, Taesang Choi, Hong-Ik Kim, Junyoung Park, and Chang Hee Yun

The authors describe a CDNI gateway model that is standard-capable and platform-independent. With the CDNI gateway model, they design and implement a complete CDNI system and conduct a CDNI service trial with three major ISPs in South Korea. According to the analysis of experimental results from the service trial, they observe that CDNI can reduce content traffic by about 40 percent at the Internet exchange (IX) link compared to a legacy CDN system.

ABSTRACT

Content delivery service has become a major traffic load on today's Internet, and this has triggered the interest of ISPs in operating their own content delivery networks (CDNs) to optimize Internet traffic considering both content delivery caching and user-network proximity. ISPs, however, are typically regionally bound or network-domain-wise isolated; hence, their CDN gain is somewhat limited. In order to enhance the gain of ISP CDN services to the level of incumbent global CDNs, a CDN interconnection (CDNI) model is introduced by IETF, where local ISP CDN services can be extended among heterogeneous CDNs across network domains. However, despite the multiple benefits of a CDNI system, it is difficult to apply a CDNI service to the current CDN market due to the platform independence. Hence, we introduced a CDNI gateway model that is standard-capable and platform-independent. With the CDNI gateway model, we design and implement a complete CDNI system and conduct a CDNI service trial with three major ISPs in South Korea. To the best of our knowledge, this is the first CDNI service trial complying with the IETF standard achieved by a multi-ISP collaboration. According to the analysis of experimental results from the service trial, we observe that CDNI can reduce content traffic by about 40 percent at the Internet exchange link compared to a legacy CDN system.

INTRODUCTION

Internet contents, such as multimedia, web pages, and application updates, has become a dominant traffic source on today's Internet. Since the growth of the content traffic volume is increasing over time, the link investment cost is increasing as well. A content delivery network (CDN) is commonly used to reduce traffic on the Internet and improve user service experience [1, 2]. In the incumbent CDN service, a CDN operator, who is independent of an Internet service provider (ISP), places CDN servers and surrogates over the world. In this model, end users receive CDN services from surrogates based on a certain level of proximity on the Internet. However, CDN servers and surrogates are typically distributed unevenly over different ISP domains. On the other hand, ISP operators are becoming interested in operating their own CDN services as they see opportunities for further improvements on the user experience and traffic control because of ownership of the networks. In turn, ISP-operated CDNs are expected to reduce a large portion of traffic in the network as more than 60 percent of the Internet traffic consists of content delivery. However, an ISP is typically regionally bound or network-domain-wise isolated, so its CDN gain is limited because its access users might receive contents from other CDNs. In order to compete against incumbent global CDNs and increase the gain of ISP CDN services, an ISP should extend local ISP CDN services across the network domains, thereby requiring interconnection between CDNs [3]. In 2011, British Telecom (BT) launched wholesale content connect (WCC), which delivers multimedia content to broadband end users from content servers situated within the BT network. A content service provider's (CSP's) content is delivered to any Internet connected end user in the United Kingdom from caches held within the BT wholesale network. ISP CDN interconnection between FT-Orange and Poland Telecom (PT) with heterogeneous solutions was tried in 2011 as well. For this, Cisco's CDS solution and Coblitz's CDN solution were used for FT-Orange and PT, respectively. Meanwhile, Cisco began a three-phase exploration into the feasibility of CDN federations, a CDN federation pilot (CFP) for testing technology interworking issues and business models [4]. Cisco defines CDN federations as multi-footprint open CDN capabilities built and shared by autonomous members. Since 2011, the Internet Engineering Task Force (IETF) has been studying how CDNs can operate in an interconnected manner under a Working Group on CDN interconnection (CDNI), envisioning a large federation of local ISPs to reduce CDN traffic between ISPs [5, 6]. IETF summarizes the three major use cases of CDNI as follows.

Footprint Extension Use Cases: This is considered a major use case. CDNI enables CDN providers who have a geographically limited footprint to provide their services beyond their own footprints. It includes use cases of geographic exten-

Yonghwan Bang, June-Koo Kevin Rhee, KyungSoo Park, Kyongchun Lim, and Giyoung Nam are with KAIST; John D. Shinn is with Solbox; Jongmin Lee and Sungmin Jo are with SK Telecom; Ja-Ryeong Koo is with LG UPLUS; Jonggyu Sung and Young-il Seo are with the Institute of Convergence Technology, KT; Taesang Choi is with ETRI; Hong-Ik Kim is with CJ HelloVision; Junyoung Park is with T-Broad Inc.; Chang Hee Yun is with the National Information-Society Agency.

0163-6804/16/$25.00 © 2016 IEEE

sion, inter-affiliates interconnection, ISP handling of third party contents, and nomadic users.

Offload Use Cases: Unexpected heavy traffic may introduce overload beyond the capability of a CDN. In this case, CDNI can offload the overloaded traffic to another CDN (downstream CDN, dCDN). Moreover, when a partial failure occurs, CDNI can redirect traffic to another CDN.

CDN Capability Use Cases: In this use case, the CDN provider may have an appropriate geographic footprint, but may wish to extend the supported range of devices and users or the supported range of delivery technologies. In this case, a CDN provider may interconnect with a CDN that offers services the CDN provider is not willing to provide, or its own CDN is not able to support.

By enabling these use cases, we can obtain many-fold benefits. Importantly, CDNI can solve inefficiencies in the current CDN server infrastructure. The number of CDN providers is increasing, and their server infrastructure is necessarily deployed redundantly. From the point of view of a network infrastructure provider, these inefficiencies can be eliminated by CDNI, and moreover it can improve the user's quality of experience. On the other hand, CDNI can save capital expenditure (CAPEX) and operational expenditure (OPEX) costs on the network infrastructure by reducing redundant content traffic especially on the Internet exchange (IX) link, for which major investment with overprovisioning is required to avoid congestion [8]. In a CDNI-enabled network, every user's content request is serviced within the user's ISP network independent from CDN contracts in order to avoid inter-ISP traffic. To reap all these benefits by interconnection to the level of those of an incumbent CDN system, we can consider the following requirements:

- First, CDNI must interconnect CDN request routers that are operated by different providers to achieve effective content request redirection across CDNs.
- Each CDN participating in CDNI service should exchange content distribution metadata through CDNI interfaces so that CSP distribution policies can be enforced consistently across CDNs.
- CDNI must provide a control protocol across CDNs so that important actions can be triggered across CDNs.
- Lastly, CDNI is required to support exchange of logging/reporting information so that essential applications such as accounting and billing, reporting, and analytics can be performed over a CDN federation.

To realize all of these system requirements, we design and implement a CDNI system based on a CDNI gateway model that is standard-capable and platform-independent. With our CDNI system developed at Korea Advanced Institute of Science and Technology (KAIST), we conducted a CDNI trial service consisting of three major nation-wide ISPs, a cable network provider, and a CSP. Three major ISPs and a cable network provider collaborated to build four CDNs, and a CSP took the role of publisher of the trial service. Through the CDNI service trial, we verified that CDNI can reduce content traffic by more than 40 percent compared to legacy CDN systems. The rest of this article is organized as follows. First, we introduce the concept and architecture of a CDNI gateway. Second, the operations for a CDNI system with the CDNI gateway model are explained. Third, the CDNI trial service is described and analyzed. Finally, we conclude with a discussion.

Architecture of CDNI Gateway

A CDNI gateway with minimum platform dependence is designed and developed with interfaces toward standalone CDNs, which requires minimum modification of a legacy CDN system. This gateway model can provide CDNI standard compliance as well as immediate legacy CDN interworking. From a technical point of view, a CDNI gateway is designed with two parts: a CDNI module and a CDN adaptation interface. A CDNI module includes a CDNI interface and a CDNI database. The CDN adaptation interface provides communication between a CDNI module and legacy CDN systems. The CDNI database manages required information for CDNI between CDNs. To optimize our CDNI system, we implement our own HTTP server and DNS server that perform parallel processing for user content requests cooperating with the CDNI system. The detailed role of each is as follows.

A CDNI Interface is an IETF standard-compliant CDNI interface application programming interface (API) to communicate with other CDN platforms for interconnection. It provides the principal functions of CDNI bootstrapping, request routing, log data acquisition, and deployment of metadata or footprint information. The CDNI interface is designed as a simple HTTP-based representational state transfer-ful (RESTful) interface. In terms of functional architecture, a CDNI interface consists of control, request routing, metadata, and logging interfaces. Each interface communicates with HTTP POST messages. Payload data is formatted in JavaScript Object Notification (JSON) [9] style. The detailed role of each interface is as follows:

- Control interface: It provides the CDNI bootstrapping process and management functions that manage the information of CDNI contracts.
- Request routing interface: If a user request is generated, a CDNI gateway HTTP server asks a request routing interface to get an appropriate dCDN based on footprint and capability information. It initializes itself with footprint information, which is provided by legacy CDN administration through a CDNI adaptation interface. It also provides a set of management functions for footprint and capability information of a CDNI network.
- Metadata interface: It provides initialization and management functions; the initialization function shares footprint and capability information at bootstrapping time with other CDNI entities, and the management functions provide ADD, DELETE, UPDATE, and other related operations of metadata.
- **Logging interface:** Every time each content request is serviced or routed, a CDNI log is saved through logging interfaces. It provides a GET method to get logging information from dedicated dCDNs or an uplink CDN (uCDN).

A CDNI database retains the information for a CDNI service: capability information of

From a technical point of view, a CDNI gateway is designed with two parts: a CDNI module and a CDN adaptation interface. A CDNI module includes a CDNI interface and a CDNI database. The CDN adaptation interface provides communication between a CDNI module and legacy CDN systems.

Category	Function	Description
Administration	RUN	Request to CDNI gateway to run CDNI module
	INIT	Send CDN information to CDNI gateway and request to initialize
	STOP	Request to CDNI gateway to stop CDNI service
Database configuration	ADD	Request to CDNI gateway to add information
	UPDATE	Request to CDNI gateway to update information
	REMOVE	Request to CDNI gateway to remove information
Get information	GET_CDmD	Request to CDNI gateway to get content distribution metadata
	GET_LOG	Request to CDNI gateway to get other CDN's (dCDN)'s CDNI log data

Table 1. Functions of the CDNI adaptation interface in a CDNI gateway.

each CDN surrogate server and footprint that describes the service scope (the allowed client IP subnet or country code, etc.). Content distribution metadata describe an authorized service timeframe or service type for each content. A CDNI service log and CDNI contract information are saved to the CDNI database as well. All information of the CDNI databases is created, updated, deleted by the CDNI interface API, and written in the JSON format.

A CDNI gateway HTTP or DNS server performs parallel processing for user HTTP requests using event queues with multi-thread processing. It accepts user HTTP/DNS requests and redirects them to final HTTP/DNS targets by asking an HTTP/DNS request routing interface of the CDNI interface. Each server can support a maximum of eight cores for which working threads are evenly scheduled. We conduct performance assessment showing that each server can process more than 20,000 content requests per second under a laboratory stress test environment.

An adaptation interface is implemented for communication between a CDNI gateway and a legacy CDN request router to minimize platform dependence. It provides API functions of the CDNI database configuration and administration to a legacy CDN. Table 1 provides detailed design information of the CDNI adaptation interface. It has three functional categories, and each category consists of HTTP method functions. A CDN administrator can boot up and stop a CDNI gateway with administration functions. Data configuration functions support database management of a CDNI gateway. When the legacy CDN needs other CDNs' information such as content distribution metadata or log data, the corresponding get-information functions can provide it.

Operations for a CDN Interconnection with a CDNI Gateway

In a CDNI system, the content delivery service is operated through a uCDN and a dCDN. When a content request is generated, a uCDN accepts the user request, and chooses an appropriate dCDN and redirects the request. The dCDN provides the content delivery service to the user on behalf of the uCDN. Since a uCDN can select itself as a dCDN, a CDNI should be dCDN-capable as well as uCDN-capable. Each CDN in a network of a CDNI system communicates only with its own CDNI gateway, and the CDNI gateway processes all CDNI-related operations. This supports CDN platform independence in a CDNI system with heterogeneous CDNs. Figure 1 shows how a CDNI system works with a CDNI gateway in detail. In a CDNI system, a CSP makes a contract of a content delivery service with a uCDN, which in turn makes another contract of CDN interconnection with multiple dCDNs. When a user requests a CSP's content service, the CDNI gateway of the uCDN accepts this request on behalf of a CDN request router, and checks the corresponding footprint information from the CDNI database to determine the appropriate dCDN among the contracted dCDNs and allow the user request redirection to the dCDN. The dCDN receives this redirected user request and performs the same process that the uCDN did, and if it finally determines there is no better dCDN than itself, it redirects the request to its CDN request router to let the CDN deliver the content to the user. The policy to determine the best dCDN must take into account various parameters including the proximity of surrogates and the user, the CDN server and surrogate loads, network bandwidth usage, and so on. In our CDNI design, we use a proximity-first policy where every content request is redirected to a dCDN that is deployed on the same carrier network with the user. This policy can reduce the IX link traffic dramatically since content delivery in a IX link only occurs when the first acquisition is acquired by the uCDN for each content service. Due to these aspects of the CDNI system, each surrogate of all CDNI participating CDNs should store a greater amount of content than a legacy CDN system. The impact of this is discussed later in this article.

In a CDNI system, a user request can traverse multiple dCDNs. Hence, the CDNI system can cause looping or flooding of request redirection messages, which can result in a service failure. It is critical to prevent looping and flooding to provide a reliable CDNI service. Looping and flooding prevention can be provided by utilizing combined information of a CDN provider ID (CPI) list and a maximum number of allowed redirection hop count (MaxNumRedHops). With this information, we implemented a looping/flooding prevention mechanism into a CDNI gateway.

1. When a CDNI gateway receives a content request or a request redirection message, it determines the next appropriate dCDN, adds its CPI information to the CPI list in the request redirection message and checks whether there are the same CPIs. If so, it is considered as a loop and chooses the second best dCDN, and so on. If there is no possible option, the request is dropped.
2. For flooding prevention, we add a NumRedHops field into the redirection message. Every time a request message is redirected, each CDNI gateway increases the value of this field by one. If NumRedHops exceeds MaxNumRedHops, which is a predefined value, the corresponding redirection message is dropped, as appears in an IETF CDNI WG draft [10].

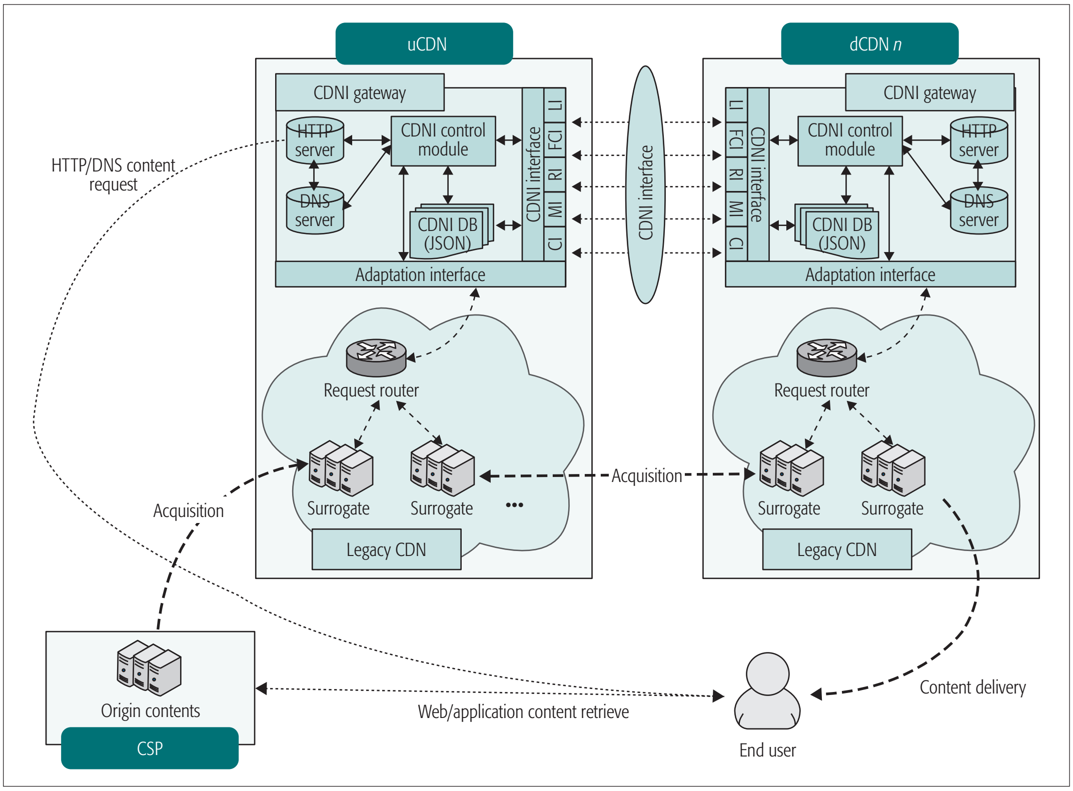


Figure 1. Service scenario of a CDNI system.

It is critical to prevent looping and flooding to provide a reliable CDNI service. Looping and flooding prevention can be provided by utilizing combined information of a CDN provider ID (CPI) list and a maximum number of allowed redirection hop count.

CDNI Trial Service

From a technical point of view, we are interested in observing the practical performance of a CDNI system in terms of traffic reduction and user experience. In particular, an IX link requires the most investment with overprovisioning to avoid traffic congestion due to large capacities of broadband core and access networks [8].

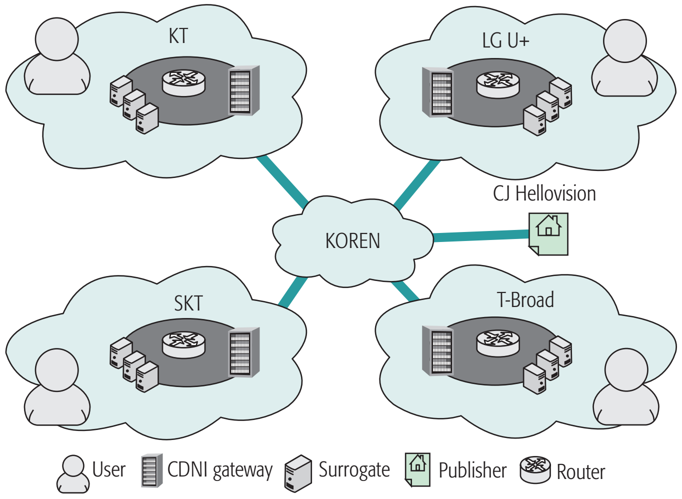
Our major concern is to verify IX link traffic reduction by application of CDNI. For this purpose, we planned a CDNI trial service and established the KAIST CDNI Consortium, which consists of three major ISPs, a cable network provider, a content service provider, and KAIST in Korea. The three major ISPs and the cable network provider collaborated to build four CDNs, and the CSP took the role of publisher in our trial service. Offered contents were provided by all CDN providers and the publisher. Each CDN interconnects with CDNI gateways, and a policy of redirection is set to proximity-first to eliminate the traffic among ISP network domains. The proximity-first policy redirects a content request to a dCDN that is located on the same carrier network as that of the user. As aforementioned, the CDNI gateway is capable of both DNS and HTTP redirection methods. We choose DNS request redirection for our service trial since it is faster than HTTP redirection.

To determine the traffic reduction due to CDNI, it is necessary to monitor and analyze an IX link. However, direct monitoring of a commercial IX link brings technical and organizational difficulties, especially for ISPs. Hence, we employed the the Korea Advanced Research Network (KOREN) to take the role of an IX link instead of using an actual commercial IX link. All service traffic generated by surrogates is delivered over an actual commercial network, while the inter-CDN traffic that corresponds to IX link traffic is delivered through KOREN links. The trial service was offered to 170 users, where approximately 100 users actively participated. A total of 121 contents with a total volume of 9.6 GB were in place for service. For a 6-day service period, 2612 content requests were generated, and a total volume of 50.5 GB of content services was delivered. The caching policy of each surrogate follows the CDN's own policy. The corresponding content and provider information is described in Fig. 2b. As shown in the figure, each content provider's number of contents and its volume varied broadly and biasedly. TBroad provides 78 percent of the total volume of contents, while SKT only provides 4 percent. This biased content service reflects a realistic content service environment.

The CDN interconnection network diagram of the CDNI trail service is presented in Fig. 2a. A trial service site that provides content download services was temporally placed. All the CDN providers collected their subscribers who volunteered for the CDNI trial service. This figure indicates that each subscriber downloaded content only from their corresponding CDN surrogates regardless of the CSP under the CDNI system.

Results and Analysis

In order to quantify the impact of CDNI on IX link traffic reduction, we analyzed the CDNI and HTTP log information in addition to a service traffic monitoring tool that monitors content service traffic at each network interface of all CDN surrogates and a publisher content server. Figure 3a shows the content popularity distribution acquired during the CDNI trial service. As shown in this figure, it closely follows a power-law distri-



(a)

Testbed	KT	LG U+	SKT	CJHV	TBroad	Total
No. of contents	7	25	15	17	57	121
Volume of contents (GB)	0.5	0.7	0.4	0.5	7.5	9.6

(b)

Figure 2. a) KAIST CDNI trail service network; b) content information.

bution with an exponential cutoff, similar to an actual service distribution (e.g., the YouTube video popularity distribution) [11], even though the counts of offered contents and users are relatively small. Since the content popularity distribution drives the service traffic pattern, and its distribution in our trial was quite close to those observed from large-scale CDN network studies, the analysis of our trial service could represent a valid observation to some extent. In a practical Internet environment, this result should be investigated more accurately under the presence of service traffic other than that of CDN. Prior to discussing the traffic reduction effect due to CDNI, we first investigated the cache-hit ratios in cases of CDNs and fully interconnected CDNIs, respectively.

As presented in Fig. 3b, the average hit ratio of the CDNI system is less than those of CDNs. Note that the CDNI system allows CDN surrogates to pull more content from the other CDN caches without increasing the storage size. In this analysis, we exclude the cache hit when the content service is delivered by a CDN surrogate located in another domain. The hit ratio of a CDNI system is degraded not because of the fewer hits but because of the more accepted content requests destined to other CDN domains. Different from others, only TBroad's CDNI hit ratio is higher than that of the CDN case since it has significantly more contents and relatively fewer domain users. Since the CDNI hit ratios of Fig. 3b include the cache hits across other network domains, we need to analyze the amount of total generated IX link traffic over a network to ascertain the impact of CDNI on IX link traffic reduction. Figure 4 represents the IX link traffic analysis results for each case of non-CDN, CDN-only, and CDNI system, respectively.

Since we analyzed service results of a real environment, the same experiment cannot be reproduced in other system models of non-CDN or CDN-only. Hence, we remanipulate CDNI service log data and create a service log for the non-CDN and CDN-only cases. Figure 4a shows the service traffic volume of each CSP content and the total of the IX link traffic. In this case, since all content requests are resolved by the CSP origin server that is located at its own network, all service traffic crosses through the IX link. This means that the total sum of the service traffic equals the total of the IX link traffic. The three plots in Fig. 4b correspond to the service-local, service-IX, and cache-IX traffic for the CDN-only system. The service-local traffic means the volume of service traffic requested by users located at the same ISP network as the uCDN. The service-IX traffic represents that of user requests from outside of the uCDN ISP domain. Finally, the cache-IX traffic is the traffic caused by content acquisition of each CDN surrogate from the origin server. For the CDN-only case, we can easily see that the total IX link traffic volume is the sum of service-IX and cache-IX volumes for each CDN. In the case of CDNI, there is no service-IX traffic since all user requests are redirected within their own local domains. Hence, the total volume of the IX link traffic of the CDNI case equals the total volume of cache-IX traffic at each CDN. Figure 4c shows the analysis results of the CDNI case. In summary, as shown in Fig. 4d, the CDN-only system

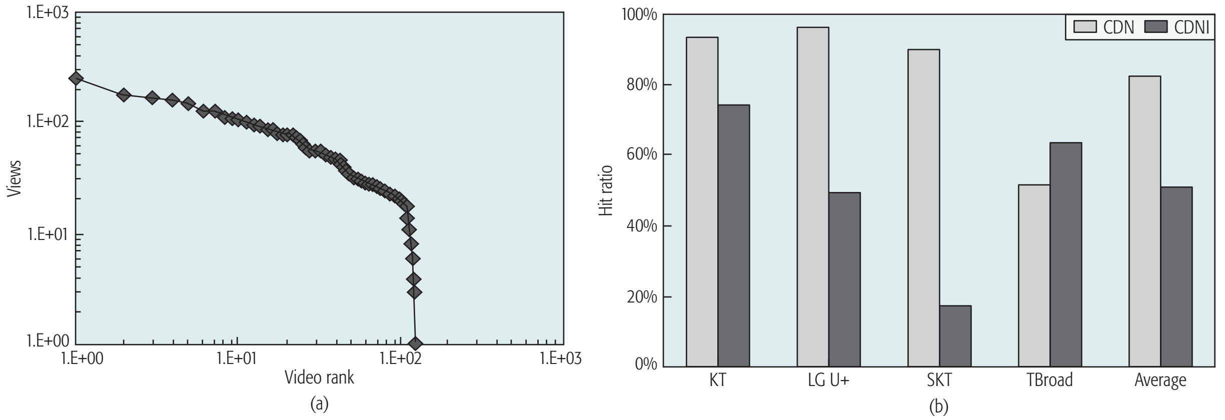


Figure 3. a) Content popularity distribution from the CDNI trial service; b) cache hit ratio at each CDN testbed.

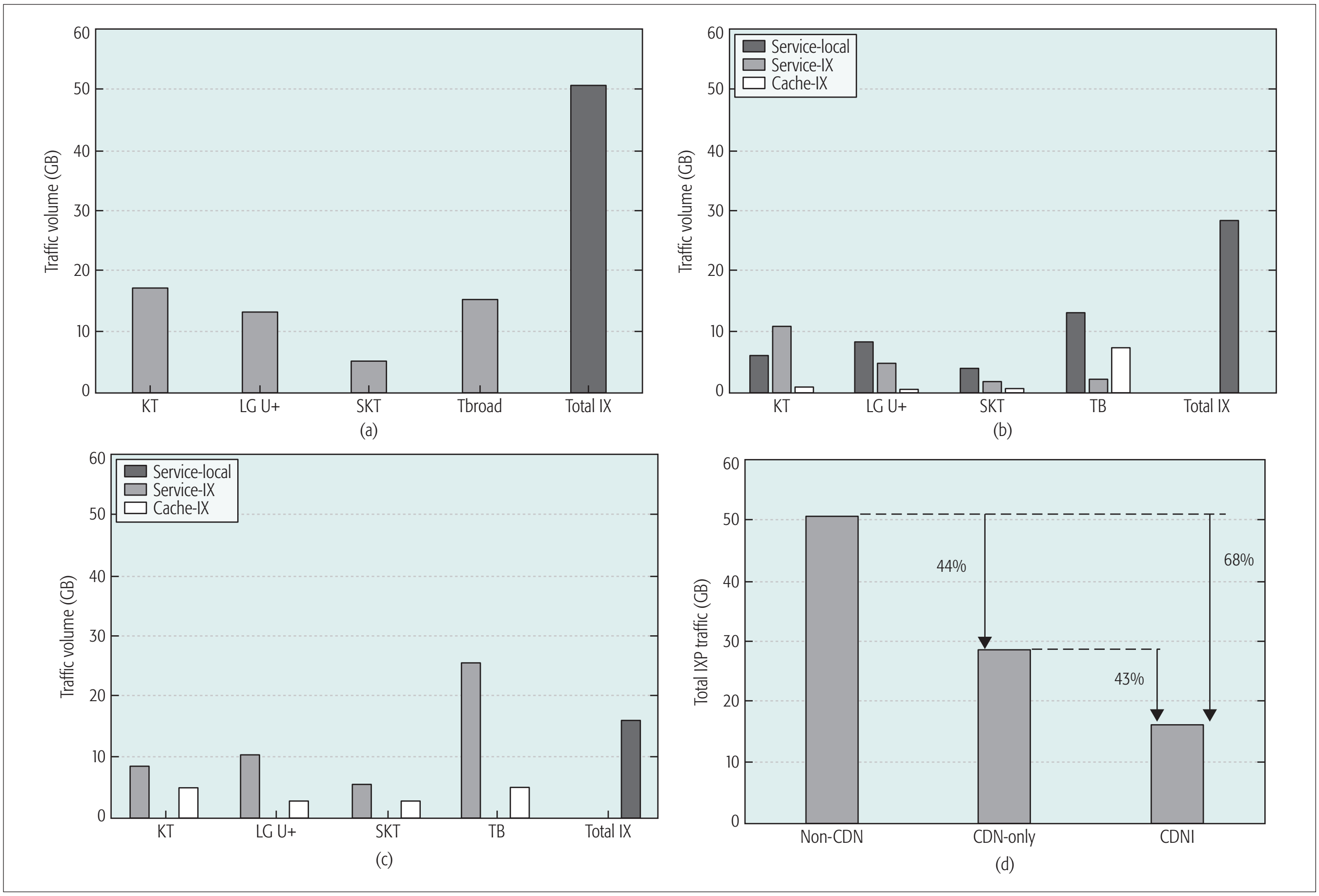


Figure 4. Traffic generation analysis results for: a) non-CDN case: service traffic volumes for CSP contents and the total volume of IX link traffic; b) CDN-only case: service traffic volumes of CDNs and the total volume of IX link traffic; c) CDNI case: service traffic volumes of CDNs and the total volume of IX link traffic; d) IX link traffic reduction ratio of CDN and CDNI systems compared to a non-CDN system.

saves IX link traffic by 44 percent compared to the non-CDN case; the CDNI system achieves a 43 percent traffic reduction compared to the CDN-only case, which is an overall 68 percent traffic reduction compared with the non-CDN case. As reviewed previously, although the CDNI system shows a lower local cache hit ratio, it eliminates much more redundant content service traffic on the IX link. This implies that the IX link traffic reduction is more important than service traffic reduction in access networks.

In addition to the observation of traffic reduction by CDNI in a wired access Internet service, we present another case study for a cellular network environment. Figure 5 shows the service throughput enhancement due to CDNI on commercial Long Term Evolution-Advanced (LTE-A) cellular networks, where two CDNs in the cellular network domains of SKT and KT are interconnected. The data show the throughput performance when SKT LTE-A mobile users request KT CDN's contents with or without CDNI. We assess the minimum, maximum, and average throughputs from 10 trials for each case. As shown in the figure, the throughput is almost doubled when CDNI is enabled. This indicates that CDNI service is not limited by IX link capacity, as expected. This result implies possible quality of experience (QoE) improvement by the measure of first-byte-play lead time through employing CDNI.

Conclusions and Discussion

In this article, we introduce and demonstrate a CDNI gateway model that provides IETF standard-compatible CDNI operations with minimum dependence on legacy CDN platforms. We verified actual traffic reduction by employment of CDNI, especially at the IX link, in a practical CDNI service trial involving three major nationwide ISPs and CSPs in Korea. From our trial results, we observe that CDNI services can reduce up to 43 percent of the content traffic volume at the IX link compared to legacy CDN systems. We proved that CDNI, especially with a gateway implementation model, is one of the most feasible practical solutions to enhance the capability of current Internet services. However, for commercial adoption of CDNI, further studies on security, billing, and business models should be carried out.

Acknowledgment

This work was supported by the ICT R&D program of MSIP/IITP, Republic of Korea [12-911-05-003, 14-911-05-001].

References

[1] B. Li *et al.*, "On the Optimal Placement of Web Proxies in the Internet," *Proc. IEEE INFOCOM*, vol. 3, Mar. 1999, pp. 1282–90.

[2] L. Qiu *et al.*, "On the Placement of Web Server Replicas," *Proc. IEEE INFOCOM*, vol. 3, Apr. 2001, pp. 1587–96.

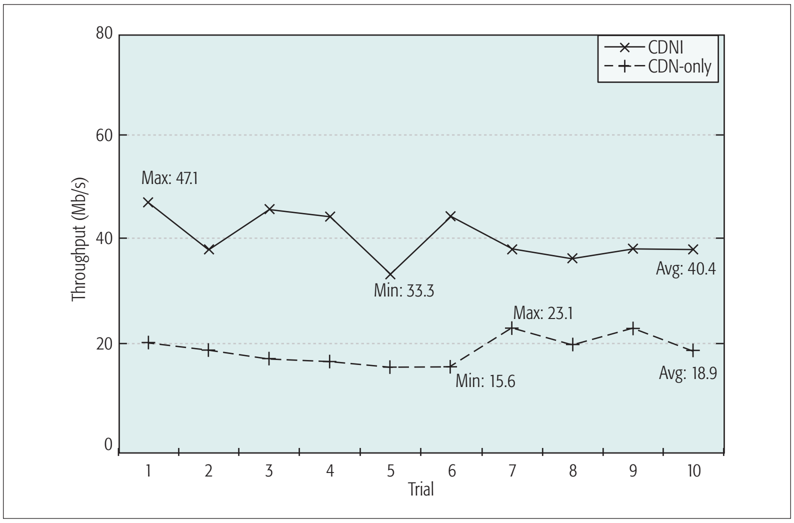


Figure 5. CDNI system performance for LTE-A cellular networks.

[3] G. Haßlinger and F. Hartleb. "Content Delivery and Caching from a Network Provider's Perspective," *Computer Networks*, vol. 55.18, Sept. 2011, pp. 3991–4006.

[4] M. Latouche *et al.*, "The CDN Federation: Solutions for ISPs and Content Providers to Scale A Great Customer Experience," Cisco Internet Business Solutions Group (IBSG), 2012.

[5] B. Niven-Jenkins, F. Le Faucheur, and N. Bitar, "RFC 6707: Content Distribution Network Interconnection (CDNI) Problem Statement," IETF CDNI WG, Sept. 2012.

[6] Y. Lee and K. Leung. "Rfc 7337: Content Distribution Network Interconnection (CDNI) Requirements," IETF CDNI WG, Aug. 2014.

[7] V. Jesus and R. L. Aguiar. "Figures of Merit for the Placement (in) Efficiency of Interconnected CDNs," *IEEE Symp. Computers and Commun.,* July 2012, pp. 277–82.

[8] C. Labovitz *et al.*, "Internet Inter-Domain Traffic," *ACM SIGCOMM Comp. Commun. Rev.*, vol 40.4., Oct. 2010, pp 75–86.

[9] T. Bray, Ed. "Rfc 7159: The JavaScript Object Notation (JSON) Data Interchange Format," IETF, Mar. 2014.

[10] T. Choi *et al.*, "Method of Request Routing Re-Direction with Loop Detection and Prevention," U.S. Patent No. 20,140,156,822. 5 June 2014.

[11] M. Cha *et al.*, "I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System," *Proc. 7th ACM SIGCOMM Conf. Internet Measurement*, Oct. 2007, pp. 1–14.

Biographies

Yonghwan Bang received an M.Sc. in electrical engineering from KAIST, Korea, in 2011 and his B.S. in computer science from Chungname National University, Korea, in 2007. He is currently a Ph.D. candidate at the Department of Electrical Engineering, KAIST. His research interests and experiences include WiFi networking, CDN networking, and system engineering. His current research interests are focused on device-to-device content delivery on 5G cellular networks.

June-Koo Kevin Rhee received a Ph.D. in electrical engineering from the University of Michigan, Ann Arbor, in 1995, and his B.S. and M.Sc. in electrical engineering from Seoul National University, Korea, in 1988 and 1990, respectively. He is currently a professor at the School of Electrical Engineering, KAIST. He has made broad contributions in the areas of optical and wireless networking, CDN networking, and optical communications with more than 120 technical journal and conference papers as author or coauthor. He has been granted more than 20 patents as inventor and co-inventor. Recently, he was awarded the Ministry Certificate of Commendation from the Korean Ministry of Science, ICT and Planning.

KyoungSoo Park is an associate professor in the Electrical Engineering Department at KAIST. He received his B.S. degree in computer science from Seoul National University in 1997, and his M.A. (in 2004) and Ph.D. (in 2007) degrees in computer science from Princeton University. After his Ph.D., he worked as an associate research scholar at Princeton University from 2007 to 2008 and worked as an assistant professor in the Computer Science Department at the University of Pittsburgh in 2009. In 2007, he co-founded CoBlitz, Inc. (acquired by Verivue, Inc. in 2010, and later acquired by Akamai,Inc. in 2012), which provides highly scalable and reliable content distribution services to ISPs and telcos.

Kyongchun Lim received his B.E. degree in electronic and electrical engineering from Sungkyunkwan University, Seoul, Korea, in 2012, and his M.S. degree in electrical engineering from KAIST in 2014, where he is currently working toward his Ph.D. degree. His current research interests include quantum key distribution and quantum information.

Giyoung Nam received his M.S. in electrical engineering from KAIST in February 2014. Previously, he received his B.S. in electrical engineering from KAIST in February 2012. He is currently at Kakao Corp. His research interests and experiences include networked computer systems, multimedia systems, network security system architecture, and network middleboxes.

John D. Shinn currently serves as a director at Solbox, a managed CDN solution provider in Korea. He has more than 16 years of platform architecture development and technical consultant experience in this industry. He has been leading the development and implementation of cloud storage platforms, cloud IaaS platforms, CDN platforms, and government IT projects in Solbox. He earned a B.S. in computer science from Yonsei University, Korea.

Jong Min Lee received his Ph.D. from KAIST in 2010. Since 2007, he has been working as an Editor of ITU-T SG13 Q3, Q8, and Q9. In 2010, he joined SK Telecom. His research area includes wired/wireless contents streaming, contents delivery optimization, and management. In 2012, he joined NGMN as an MCDO project leader and is also involved in MPEG as a WG Chair and Editor. He has over a decade's experience in information and telecommunication over fixed and mobile IP streaming technology.

Sungmin Jo received his B.S. and M.S. degrees in electrical engineering from Kookmin University, Seoul, Korea, in 1994 and 1996, respectively. Since 1996, he has been with Network Technology R&D center, SK Telecom. His research areas of interest include wired/wireless convergence networks, optical and RF relay stations, wireless LAN networks, and CDNs. He was the recipient of the Best Practice in Multimedia Innovation at the IEEE International Conference on Multimedia and Expo 2015.

Ja-Ryeong Koo received an M.Sc. in computer science from Texas A&M University at College Station in 2008 and a B.S. in computer engineering from Kyung-Hee University, Korea, in 2003. He is currently a deputy manager at LG UPLUS. He has been working on the CDN platform and networking for IPTV broadcasting, including mobile IPTV, particularly the video portal.

Jonggyu Sung received B.S. and M.S. degrees in electronic engineering from the KyungPook National University, Korea, in 1994 and 1996, respectively. He is currently a principal research engineer at Infra Lab, Institute of Convergence Technology, KT. He has made contributions in the areas of network management and contents delivery networking with technical journal and conference papers as author or coauthor. His current research interests include network virtual functions, software defined networks, cloud computing, and next generation operations support platforms.

Young-il Seo has more than 20 years' extensive experiences as an IP network engineer at KT Network R&D Laboratory. As a key accomplishment, he successfully deployed KT NGN and implemented KT TPS, including IPTV over KT NGN. He is responsible for design, deployment, and engineering of KT's IP network. He was the Editor of ITU-T IPTV FG and is now active in the IETF P2P related WG. He is focusing on next generation content delivery technology, P2P issues, and IETF application-layer traffic optimization technology.

Taesang Choi is a principal member of engineering staff in ETRI, having joined the Institute in 1996 after doing R&D on network and service management of telecommunications during his Ph.D. studies at the University of Missouri at Kansas City. He has successfully managed a number of projects in the area of networking, especially in Internet traffic engineering, measurement and analysis, QoS, SDN, and NFV management. Currently, he is managing the OPEN-TAM subproject in the ONOS SDN Open Source consortium, and is involved in a Korean government supported project, "Smart Networking Core Technology Development," which addresses SDN/T-SDN/NFV control and management for carrier-grade networks.

Hong-Ik Kim received his B.S. degree from the Department of Electronics at Hankuk Aviation University, Gyeonggi, Korea, in 1996 and his M.S. and Ph.D. degrees from the Department of Electrical and Communication Engineering at Hanyang University, Seoul, Korea, in 2003 and 2007, respectively. Currently, he is a principal researcher at CJ HelloVision. His current research interests include pattern recognition, multimedia on-demand systems, platform convergence, broadband communication, and data broadcasting.

Junyoung Park received an M.S. in telecommunication engineering from the University of Yeungnam in 1996. He has 15 years' work experience in broadcast and telecommunication engineering in the cableTV industry. He is currently a team leader of the technology strategy team in T-Broad Inc., South Korea.

Chang Hee Yun earned an Honors Bachelor of Management Information Systems dein DanKook Yniversity, Korea, and graduated at the top of his Master's class at the university in 1999. He has also completed the Master's course at the University for Peace in 2014 as well. He is currently working as a director at the National Information-Society Agency, Korea. He has worked as a project manager in building for network infrastructure of the ITU Plenipotentiary 2014, Busan, Korea. At that time, he planned and coordinated the wireless and wired network mobile application to operate during the whole conference, which was successful, resulting in paperless operation for the entire conference. He earned an Achievement Award in 2015 for successfully hosting ITU Plenary 2014 from the Ministry of Science, ICT, and Future Planning.

Call for Papers
IEEE Communications Magazine
Internet of Things (IoT)

Background

Internet of Things is seen as a set of vertical application domains that share a limited number of common basic functionalities (such as communications and networking protocols and operating systems APIs). In this view, consumer centric solutions, platforms, data management, and business models have to be developed and consolidated in order to deploy effective solutions in the specific fields. The availability of low cost general purpose processing and storage systems with sensing/actuation capabilities (now available also to prosumers) coupled with communication capabilities are broadening the possibilities of IoT leading to open systems that will be highly programmable, virtualized and will support large numbers of APIs. Internet of Things emerges as a set of integrated technologies new exciting solutions and services that are set to change the way people live, produce goods. Internet of Things is rewarded by many as a fruitful technological sector in order to generate revenues. IoT covers a large wealth of consumer centric technologies (from sensors to communications up to software platforms) and it is applicable to an even larger set of application domains (from manufacturing to e-health, from logistics to automotive). Innovation will be nurtured and driven by the possibilities offered by the combination of increased technological capabilities, new business models and the rise of new ecosystems.

This proposed Feature Topic (FT) issue will gather articles from a wide range of perspectives in different industrial and research communities of IoT. The primary FT goals are to advance the understanding of the challenges faced in IoT communications, networking, distributed processing, new signal processing capabilities, software platforms and end – users devices over the next decade, and provide further awareness in the IoT research communities on these challenges, thus fostering future investigation. In addition a perspective on the business possibilities of IoT are of interest in order to enable and deploy the foreseen technical solutions. Original research papers are to be solicited in topics including, but not limited to, the following themes:

- Existing and future communication architectures and technologies for large IoT systems
- Existing and future use cases and deployment of large IoT systems
- Design and evaluation of large IoT test beds, prototypes, and platforms for consumer centric IoT application development and deployment
- Identification of viable business models and related ecosystems
- Solution and services supported by consumer devices
- Security, Privacy and interworking issues for cooperative IoT operations
- Interfaces, cross-platform communication and programmability for IoT systems
- Autonomics mechanisms for QoS and performance evaluation for IoT solutions
- Game-theoretic and control-theoretic mechanisms for IoT resource allocation and management
- Integrating 4G and 5G wireless technologies into IoT communications and Platforms
- Integration of cognitive techniques with IoT systems
- Energy-efficient communications considering opportunistic policies for large IoT systems
- Big data and data analytics solutions for IoT systems
- Comparison and improvement of IoT communication protocols
- Novel distributed techniques (e.g., Edge/Fog computing)
- New sensing and actuation capabilities and devices and their applicability

Submissions

Articles should be tutorial in nature, with the intended audience being all members of the IoT research community. They should be written in a style comprehensible to readers outside the specialty of the article. Mathematical equations should not be used (in justified cases up to three simple equations are allowed). Articles should not exceed 4500 words (from introduction through conclusions). Figures and tables should be limited to a combined total of six. The number of references is recommended not to exceed 15. In some rare cases, more mathematical equations, figures, and tables may be allowed if well justified. In general, however, mathematics should be avoided; instead, references to papers containing the relevant mathematics should be provided. Complete guidelines for preparation of the manuscripts are posted at http://www.comsoc.org/commag/paper-submission-guidelines. Please send a pdf (preferred) or MSWORD formatted paper via Manuscript Central (http://mc.manuscriptcentral.com/commag-ieee). Register or log in, and go to Author Center. Follow the instructions there. Select "December 2016/IoT" as the Feature Topic category for your submission.

Important Dates

- Submission Deadline: June 15, 2016
- Notification Due Date: August 15, 2016
- Final Version Due Date: September 15, 2016
- Feature Topic Publication Date: December, 2016

Guest Editors

Roberto Minerva
TIM Lab, Italy
roberto.minerva@telecomitalia.it

Mohsen Guizani
University of Idaho, USA
mguizani@uidaho.edu

Christos Verikoukis
CTTC, Spain
cveri@cttc.es

Hausi Muller
University of Victoria, Canada
hausi@cs.uvic.ca

Soumya Kanti Datta
EURECOM, France
dattas@eurecom.fr

Yen-Kuang Chen
INTEL, USA
y.k.chen@ieee.org

Computing for Rural Empowerment: Enabled by Last-Mile Telecommunications

Somen Nandi, Saigopal Thota, Avishek Nag, Sw. Divyasukhananda, Partha Goswami, Ashwin Aravindakshan, Raymond Rodriguez, and Biswanath Mukherjee

The authors aim to guide service providers, industry practitioners, and local entrepreneurs with a technology-and-deployment-trend analysis to choose, deploy, and operate suitable telecommunication networks depending on the unique features of the rural/remote area. Their goal is to bring attention to accessible and affordable technologies with practical considerations.

Abstract

Increasing economic and educational exposure, and promotion of global health and wellness can be achieved through the power of sharing knowledge, technology, and resources. ICT can play a key role in disseminating such knowledge across the world. But a digital divide exists between urban and rural/remote areas, which results in economic and social disparities across regions. Developing last-mile telecommunication technologies for rural/remote areas is a crucial aspect in providing computing and ICT services that can integrate millions of stakeholders in rural/remote areas globally into the digital age, particularly with the advent of cloud computing. This article focuses on the different aspects of providing last-mile rural telecommunication access such as interfering factors, technology options, and deployment trends. This article aims to guide service providers, industry practitioners, and local entrepreneurs with a technology-and-deployment-trend analysis to choose, deploy, and operate suitable telecommunication networks depending on the unique features of the rural/remote area. Our goal is to bring attention to accessible and affordable technologies with practical considerations.

Introduction

Computing for Rural Empowerment

It has been well studied that increasing global economic citizenship and educational exposure, and promotion of global health and wellness, can be achieved through the power of sharing knowledge, technology, and resources. Computing and information and communication technology (ICT) services can play an important role in disseminating such knowledge around the world. Computing technologies and services can potentially reach large populations faster in emerging economies. Especially with the advent of the cloud-computing paradigm, there is new scope for rural users to exploit today's technologies such as cloud computing and cloud storage, where the resources required and the charges at the user site to access these services are minimal using devices such as smartphones. Combining the power of cloud computing along with existing or emerging web services will pave the way for bridging the digital gap and can bootstrap rural economy and the quality of life [1].

Our goal is to bring attention to accessible and affordable technologies with practical considerations. In this article, we discuss various technology options, deployment trends, and best practices in leveraging the power of ICT, particularly for rural empowerment. The interfering factors with possible resources integration [2] to utilize the existing infrastructures for last-mile communications in rural areas are also discussed.

Last-Mile Rural Telecommunication Networks

Telecommunication networks play a crucial role in connecting rural users to the cloud and the rest of the Internet [3]. However, rural/remote areas are characteristically influenced by factors such as scattered user base, resistance to adopt new technology, and affordability. These factors result in limited or nonexistent last-mile connectivity infrastructure. They also create a digital divide between urban and rural/remote locations, which results in lack of access to computing infrastructure, and leads to economic and social disparities across regions. Revolutionary technologies and appropriate implementation plans need to be explored for rural telecommunication networks, which should be robust, flexible, scalable, affordable, and easy to use. We also believe that isolated efforts and investments in one area without integration among multiple areas will likely be unsuccessful.

Typically, a telecommunication network consists of three major infrastructure categories:

- Access network
- Metropolitan area network
- Core network (Fig. 1)

The access network enables end users (businesses and residential customers) to connect to the rest of the network, and it typically spans a few kilometers. The network that connects the access network to the rest of the network is generally referred to as a backhaul/backbone/backend network. It consists of the metropolitan and core networks. The metropolitan network spans a metropolitan region, covering distances of a few tens to a few hundreds of kilometers. A metro-core aggregation ring, as the name suggests, aggregates traffic from multiple metropolitan access networks using aggregation switches and routes the traffic to the core network. The core net-

Somen Nandi, Saigopal Thota, Avishek Nag, Ashwin Aravindakshan, Raymond Rodriguez, and Biswanath Mukherjee are with the University of California, Davis; Sw. Divyasukhananda is with Ramakrishna Mission Vivekananda University; Partha Goswami is with the Indian Institute of Technology, Kharagpur.

0163-6804/16/$25.00 © 2016 IEEE

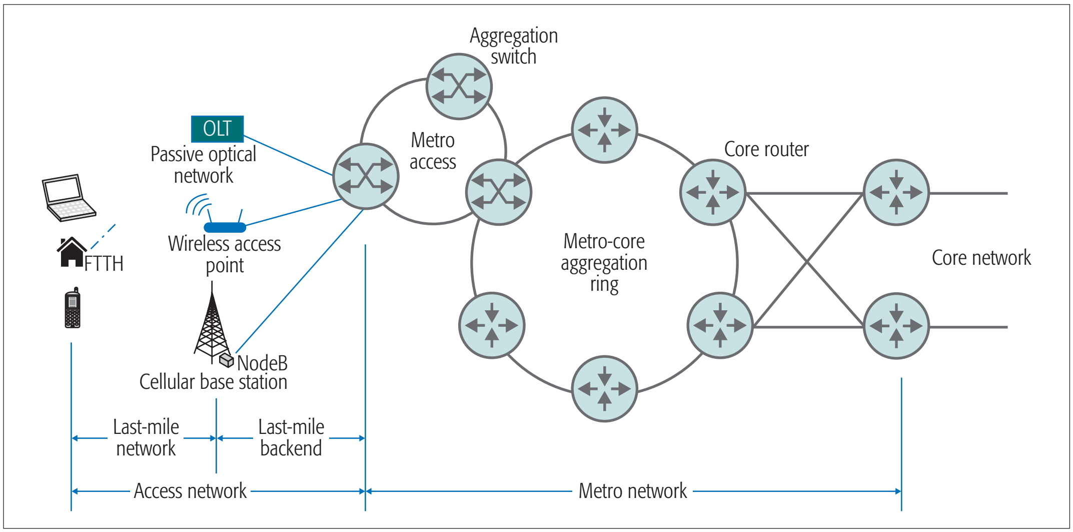


Figure 1. Major infrastructure categories in a telecommunication network.

Affordability of a service by end-users is the ultimate driving force for developing a financially sustainable solution. Therefore, low-cost network solutions are required for rural areas. Customer density and economic conditions of users determine the selection of a technology.

work provides global connectivity with the help of core routers and switches, and spans long distances such as a few hundred to a few thousand kilometers. Network operators, service providers, and researchers continue to address challenging issues to accommodate higher bandwidth requirements with increasing traffic. But an important challenge yet to be addressed is how to provide cost-effective last-mile connectivity to rural areas for enabling computing services.

Our detailed economic analysis on cell phone and Internet penetration with respect to income levels of 50 nations [4] shows that the Internet penetration today is tightly correlated to average income levels, with only 31 percent penetration in developing nations. Cell phone penetration, on the other hand, is widely observed throughout the world with 96 percent global penetration (89 percent in developing nations). This is a strong example showing how a technology penetrates into a society, and people will adopt it, irrespective of their economic status, if it is made affordable and easy to use with compelling applications. Smartphones can play a major role in this context as they provide both cellular and Internet connectivity. Even though smartphones are not yet used widely in rural areas, their penetration will increase in the next few years similar to cell phones, especially with their prices going down [5].

Integrated Approach

Most previous works on last-mile rural computing and communication were reported before the recent cloud computing and smartphone era. In this article, we present an integrated approach from technological, business, and sustainability perspectives — as well as available resources — for providing computing services with support from telecommunications to improve rural living standards.

In this article, we discuss constraints and interfering factors for developing computing services and telecommunication networks in rural areas and technology options. Then we provide a comprehensive review of relevant research activities during the recent past to help researchers, stakeholders, and industry:

- To gain knowledge about existing solutions around the world
- To further improve the state of the art in rural telecommunications research
- To integrate technologies and network solutions for deployment in diverse regions of the world

Our goal in this article is not only to provide a comprehensive review of multiple technologies and deployment trends around the world, but also to elucidate the sustainability of services in a rural setting that needs a multidisciplinary integrated approach.

Interfering Factors

Several factors are crucial in determining a viable technology solution for last-mile connectivity in rural scenarios, as reviewed below.

Geographic Location

The geographic location determines the terrain and hence the challenges associated with its characteristics (i.e., flat land, hilly areas, dense forest areas, etc.). A hilly and densely forested area may have more fading (i.e., reduced signal intensity due to the propagation of a signal over multiple paths and interference) and signal power loss compared to a relatively flat area with less tree canopy [6]. Location governs the cost associated with the infrastructure development and transportation of telecommunication equipment, troubleshooting, and maintenance of the network.

Economic Conditions

Affordability of a service to end users is the ultimate driving force for developing a financially sustainable solution. Therefore, low-cost network solutions are required for rural areas. Customer density and economic conditions of users determine the selection of a technology. For example, WiFi operates on unlicensed industrial, scientific, and medical (ISM) spectrum, which is open for usage without a regulatory fee, whereas WiMAX and Long Term Evolution (LTE) require licensed spectrum (which comes with a spectrum-usage license fee that will be billed to users indirectly). Therefore, a suitable network solution needs to be selected based on the affordability of users with appropriate return on investment (RoI) to the service provider.

Deployment trend/ technology	Cost		Coverage/ penetration	Lic./ unlic.	NLOS*	Data rate	Power	Deterioration/ interference	Remarks
	Infra.	Oper.							
Long-reach WiFi	Low	Low	Low	Unlic.	No	Low	Low	High	Uses off-the-shelf equipment, hence low cost
WiMAX-based	High	High	High	Lic.	Yes	High	High	Low	User devices do not have WiMAX interfaces, need WiFi last hop
Delay-tolerant networks	Low	Low	High	Unlic.	Yes	Low	Low	Low	Delayed response time, not suitable for real-time applications
Hybrid wired and wireless	High	Med.	Med.	Unlic.	Yes	High	Low	Med.	Low cost and flexible connectivity
Cellular	High	High	High	Lic.	Yes	Med.	High	Med.	Provides both voice and data communication; rapidly penetrating technology
Cognitive radio	High	Low	High	Lic.	Yes	High	Low	High	Can exploit unused licensed spectrum
Power line communications	Med.	Low	High	Unlic.	Yes	High	Low	High	High penetration of power lines is an advantage
MIMO wireless	High	Low	High	Both	Yes	High	Low	Low	Expensive
Alternative telecom networks	Low	Low	High	Both	Yes	High	Low	N/A	Leverage existing infrastructure, not widely available

*NLOS: non-line of sight; Lic.: licensed; Unlic.: unlicensed.

Table 1. Deployment trends based on different network technologies: pros and cons in the context of last-mile rural connectivity.

Motivation/Incentives and Adoptability

In developing economies, there are many remote but populated regions with barriers to development such as lack of transportation infrastructure that contribute to economic and knowledge disparities, resulting in lack of entrepreneurship and social innovation. Insufficient knowledge among these bottom-of-the-pyramid customers may lead to ignoring and underestimating the benefit and power of ICT. Community-based participatory research (CBPR) is essential to understand and motivate end users to become integral partners in this multi-stakeholder value chain. It is important to study and analyze cultural and behavioral incentives in rural areas to improve adoption of technology and make people act in a certain way, customized to the prevailing culture. In order to have an effective CBPR in developing or rural regions, it is unavoidable to implement any program without the direct engagement of the local political and cultural (and often religious) leaders. Thus, any multi-faceted project should always consider involving both political and cultural constraints along with other interfering factors.

A single "one size fits all" solution does not exist for telecommunication in rural areas. Based on different interfering factors presented here, a network solution needs to be selected for a particular scenario.

Sustainable Business Framework

Rural areas are typically characterized by lack of skilled personnel, sparse population distribution, difficult terrains for transporting equipment, and so on. A common observation in rural computing projects is that they are not managed/maintained after the group who set up the project leaves the site. Sometimes, projects are discontinued due to lack of funds to run the infrastructure.

Therefore, community-based network connectivity projects for rural computing need to be designed to be autonomous and remotely manageable, and require minimal human intervention. The cost of operation should be low with minimal dependence on external resources such as electricity supplied from distant power stations. The fewer the dependencies, the easier the deployment, and the higher the sustainability of the project.

Strong business models make a project/investment profitable and sustainable. A recent study concluded that new technology adoptions and diffusion models are needed for rapidly evolving mobile technologies [7]. Self-sustaining business models are needed so that the infrastructures are owned and managed by local entrepreneurs besides service providers (see [8]).

Technology Options

Technology options for last-mile rural access along with their advantages and limitations are summarized in Table 1 and described below. For details, please see [5].

Wired

Wired technologies include copper or fiber-based telecommunication technologies. They provide higher data rates, and, unlike wireless technologies, they are less susceptible to external factors such as interference, signal loss, and line-of-sight (LOS) requirements. Depending on whether there is a wired infrastructure nearby, deploying last-mile connectivity using digital subscriber line (xDSL) technologies or a passive optical network (PON) with fiber to the home/curb (FTTH/C) is a viable option. In FTTH/C, access network connectivity is provided using optical fibers to the home or curb, respectively.

Fixed Wireless

Fixed wireless broadband refers to technologies where customer premises equipment (CPE) at a user's site connects to a wireless network. They include very small aperture terminal (VSAT), IEEE 802.11 (WiFi), and IEEE 802.16 (WiMAX) technologies. In VSAT, a small satellite transmitter and receiver communicate with VSAT access satellites. VSAT could be a good candidate for setting up broadband links in remote locations. WiFi provides short-range communication with speeds of up to 54 Mb/s today and throughputs of around 500 Mb/s using emerging standards, such as IEEE 802.11ac.

For providing last-mile broadband access in rural areas, most deployments observed today are based on wireless technologies due to their cost effectiveness, flexibility, and ease of installation, especially in challenging terrains. However, the backhaul connection from the wireless terminal to the core network generally uses wired (fiber or copper) connectivity.

Deployment Trends

A viable deployment trend with an appropriate business model is necessary for sustainability. A study has shown multiple benefits of improving mobile access and smartphone use; for example, Safaricom in Kenya developed a mobile payment platform named M-Pesa, where airtime can be used as currency. An Android-based mHealth system, implemented in Kenya to monitor and perform clinical care (for about two million people) during home visits in resource-constrained environments, turned out to be a viable and cost-effective solution at scale to collect electronic data during household visits [14]. We have summarized multiple deployment trends (based on technology options) for last-mile connectivity in Table 2; these are described below. Due to limitations on space and number of references, further information can be found in our detailed technical report [5], including citations to work not cited directly in Table 2.

Long-Reach WiFi

Research and deployment in setting up long-reach WiFi networks is quite mature due to the commercialization of WiFi, availability of low-cost off-the-shelf equipment, and WiFi's operation in unlicensed frequency spectrum.

Deploying WiFi-based long-distance networks has been explored with links as long as 50–100 km. Real-world deployments give poor end-to-end performance as the IEEE 802.11 medium access control (MAC) protocol is developed for short-range communication. To overcome the shortcomings of WiFi over long distances, essential changes are proposed such as an adaptive loss recovery mechanism, showing 2–5-fold improvement in Transmission Control Protocol (TCP)/User Datagram Protocol (UDP) throughput, and time-division multiple access (TDMA) on frequency bands with high signal loss during transmission.

In a different work, a testbed is evaluated on two links in Africa — Merida to El Baul (279 km) and El Aguila to Platillon (382 km) [6] — to measure the performance of very-long-distance single WiFi links. For such extremely long distances, an unobstructed line of sight and at least 60 percent of the first Fresnel zone are required. The data rates were around 65 kb/s with unmodified IEEE 802.11 protocol. By modifying the 802.11 MAC protocol to TDMA, the throughput increased to 600 kb/s, allowing video transmissions.

In [9], a long-range WiFi network is proposed with relay nodes between the end-user terminal and a rural telecenter. WiFi relay points are solar-powered self-sustainable units with their own omnidirectional antenna module. The nature of power consumption of the equipment was optimized to suit the solar power supply system.

Cellular Networks

Cellular penetration and tele-density in rural areas is significant in most countries; hence, cellular networks can be an efficient last-mile solution for rural areas. They have the advantage of providing voice as well as data connectivity, which can enable multiple services, particularly with the advent of cloud computing and smartphones.

The lack of infrastructure is pronounced in sub-Saharan Africa, making today's cellular approaches challenging to deploy. Unlike the majority of African villages, Macha and Dwesa host local wireless networks through satellite gateways. Cell phones are more prevalent and easier to use than PCs; therefore, a low-cost GSM system called VillageCell is developed to provide localized cellular coverage integrating voice over IP (VoIP) in a cost-effective manner [5]. The work presents a software-defined-radio (SDR)-controlled software implementation of the GSM stack, called OpenBTS, where core cellular services are provided for a fraction of the cost of a commercial base station, offering local cellular coverage and standard phone connections to callers using off-the-shelf equipment.

Another solution for cellular connectivity in sparse rural areas is to use a small low-power cellular base station, called a femtocell, which connects to a broadband network and provides cellular connectivity within its coverage range. Femtocells can support 16–64 simultaneous calls covering a radius of 1.5 km. Providing cellular connectivity using femtocell reduces the cost from $200,000 (for a macrocell, i.e., a traditional cellular base station) to $100 (for a femtocell). In [10], an analysis of a long-reach WiFi backhaul to support femtocells shows that a large number of simultaneous high-quality voice calls can be supported with the solution, and the number varies w.r.t. remoteness of the femtocell.

WiMAX and LTE

WiMAX-based solutions are important for last-mile connectivity in rural areas as WiMAX has greater coverage and supports broadband applications in LOS and non-LOS (NLOS) scenarios. Where no infrastructure exists, WiMAX is cheaper and faster in getting a large area covered, so it is a potential "greenfield" solution. We proposed and deployed a network infrastructure to provide education and healthcare services in India where WiMAX base stations provide blanket network coverage in rural areas, and outdoor CPE is used to connect to the WiMAX base station (Fig. 2) [1]. We observed that video conferencing and related applications can be provided in rural areas for low cost with bandwidths as low as 301 kb/s.

Cellular penetration and tele-density in rural areas is significant in most countries and hence cellular networks can be an efficient last-mile solution for rural areas. They have the advantage of providing voice as well as data connectivity which can enable multiple services, particularly with the advent of cloud computing and smartphones.

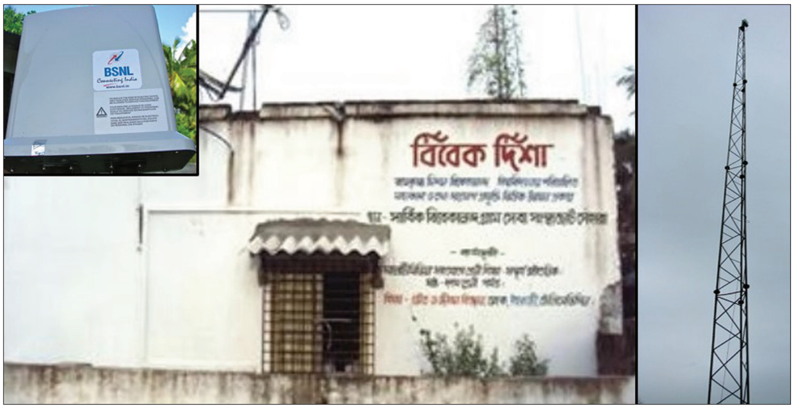


Figure 2. The education and healthcare center equipped with our WiMAX-based last-mile telecommunication solution. The figure on the right shows the mast used to host the WiMAX CPE, and the inset shows the WiMAX CPE that connects to the base station [1].

A local-loop WiMAX access network is deployed in the Siyakhula Living Laboratory, a joint venture between Rhodes University and the University of Fort Hare, for introducing ICT in rural areas that are home for 42.5 percent of the total population of South Africa, while fixed-line density in some rural areas is less than 5 percent. Reference [5] reviews a work that configures local distributed access nodes (DANs) (client systems running Ubuntu) placed in schools. Wireless access points (APs) at DANs provide access to users. Alvarion BreezeMAX technology is used for WiMAX deployment. A WiMAX micro base station is housed at the highest point in a school. DANs connect to the Internet through the base station over a VSAT connection provided by Telkom, a telecommunications provider in South Africa. The work presents a real deployment of a distributed architecture using low-cost equipment.

Long Term Evolution (LTE) is a fourth generation (4G) cellular technology, which is compatible with previous mobile technologies GSM, GPRS, UMTS, EDGE, CDMA2000, and so on. LTE allows very high user mobility (up to 450 km/h) and can extend the battery life of mobile terminals. But deployment of a WiMAX network is much cheaper than deployment of an LTE network, so it is a great choice for private mobile broadband wireless networks, as the above examples indicate.

Delay-Tolerant Networks

Low-cost connectivity alternatives such as store-and-forward networking are suitable for rural areas with minimum or no existing infrastructure and where basic data communication is more important than time sensitivity. Applications such as browsing can be delay-tolerant, where the response to a data request/query (e.g., information, documents, videos, email) will be received in a few hours after the query is submitted.

One such delay-tolerant networking (DTN) solution is deployed in a rural area with about 1000 customers where WiFi APs are set up, and the other end of the WiFi connection is in moving public transport vehicles that frequently shuttle between an urban area and a rural area. WiFi APs cache data requests (queries) made by users, and these vehicles collect queries from APs within the range of the traveled routes. The queries are sent to and relevant content is downloaded from the Internet when the vehicles are within range of Internet-enabled hotspots (which are coverage areas of wireless APs with Internet connectivity) in the urban area. This technology is inexpensive with approximate costs of $0.03 per capita, and it has been successfully tested for rural connectivity in Cambodia and India [5]. But these networks may incur a large response time (between query submission and reception of response) dictated by the frequency of the moving vehicles.

Cognitive Usage of Unutilized Television Spectrum

Cognitive radio technology enables utilization of unused licensed spectrum by sensing the environment and adapting accordingly. The IEEE 802.22 wireless regional area network (WRAN) standard is based on opportunistic usage of very-high-frequency/ultra-high-frequency (VHF/UHF) TV bands, called TV white space (TVWS).

Techniques and deployment scenarios are presented in a work (reviewed in [5]) with cognitive usage of TVWS as a possible solution for last-mile rural connectivity. They can provide wireless broadband access to rural and suburban areas with an average coverage radius of 33 km (and up to 100 km). Large network coverage and availability of white space in the spectrum make this technology particularly suitable for rural deployment. The cost-demand mismatch in rural areas is solved due to free usage of licensed spectrum and large coverage.

A wireless broadband access network, called Hopscotch, was deployed on the west coast of Scotland where point-to-point (P2P) links and blanket coverage similar to WiFi, but using white spaces in UHF, are used to provide network coverage [12]. The advantages of using UHF are wider coverage and non-LOS links. P2P links are used for backhaul, and point-to-multipoint (P2MP) links are used for providing blanket coverage. WiFi in 5 GHz is used in conjunction to serve subscribers in close vicinity. Substantial reduction in path loss and improved throughput are observed at UHF frequencies compared to 5 GHz, especially in longer NLOS links [12].

Power Line Communication

Power line communication (PLC) is another solution for broadband access. It enables utility companies to deploy communication networks and transmit data signals over existing power line infrastructure. Electromagnetic waves carrying information-bearing signals propagate via the medium voltage (MV) and low voltage (LV) lines, together with electric power. High-speed transmission of data, voice, video, and so on via power cables would be invaluable for rural areas as the electricity infrastructure generally reaches most rural areas, thereby reducing the telecommunication capital expenditure. A PLC model for broadband over power lines with MV or LV nodes converting IP-based communication signal to other suitable signal for transmission through power lines is proposed in [13]. The work gives a detailed account on the components required for a PLC system.

A company called Xeline conducted trials

Last-mile technology	Access backend	Article	Antenna	Remarks
WiFi	WiFi	B. Raman *et al.* R. Patra *et al.* R. Flickinger *et al.* [6] K. Ab-Hamid [9]	Directional Directional Directional Directional	Long-distance multihop WiFi links as backhaul for connecting multiple villages Long-distance WiFi links with adaptive loss-recovery mechanism showing improvements in TCP/UDP throughput and TDMA. Very-long-distance links tested with endpoints on hilly areas for LOS. Link distances of up to 382 km. Multihop network with relays powered by solar, making network self-sustainable in hilly terrain.
Cellular	Satellite	A. Anand *et al.*	OpenBTS implementation for GSM	VoIP-based calling as using cell phones is easier than PCs in hilly areas.
Cellular	WiFi	A. Dhanunjay *et al.* J. Fitzpatrick [10]	OpenBTS-based GSM microcells and wireless mesh Femtocells	Providing Internet and cellular voice services using solar-powered low-power microcells. In case of low user density, femtocells are a good alternative to provide cellular connectivity with a long-distance WiFi backhaul.
Cellular	Any	F. Simba *et al.*	3G omnidirectional	Discusses advantages of UMTS 900 MHz for lower path loss and higher coverage at the expense of lower bandwidth.
WiMAX	Satellite	I. Siebrger *et al.*	WiMAX local loop	WiMAX local loop created using Alvarion BreezeMAX technology and connected to the Internet using VSAT.
WiFi	WiMAX	P. Goswami *et al.* [1] D. Chieng *et al.*	Omnidirectional WiMAX base station Omnidirectional WiMAX base station	WiMAX base station provides blanket coverage, and outdoor CPE is used to connect to the base station and WiFi for local access. Multi-tier multihop WiMAX and WiFi backhaul and WiFi access tiers with a WiMAX base station serving multiple hexagonal cells with WiFi routers.
WiFi	Moving vehicles	A. A. Hasson S. Issacman *et al.*	Mobile access points Mobile access points	Requests queried are stored and forwarded using public transport vehicles. Data transferred from an area with Internet connectivity through vehicles. Similar to above, but uses low-bandwidth cellular data for sending user queries to reduce response time.
WiFi	Optical fiber	S. Sarkar *et al.* [11]	Omnidirectional mesh	Combines the high-speed connectivity of fiber and the flexibility and cost effectiveness of wireless.
TVWS	Cellular	A. Achtzehn	Cellular omnidirectional antenna	Uses TVWS for greater cellular bandwidth on existing cellular towers.
TVWS	Any	C. McGuire *et al.* [12]	Omnidirectional	Leverages greater coverage area of spectrum in TVWS to provide blanket coverage.
Power lines	Power lines	A. M. M. Altrad *et al.* [13]	Not applicable	Uses existing power line distribution infrastructure to homes to provide network connectivity.
Wireless (WiFi)	Power lines	D. Fink *et al.* A. M. Sarafi *et al.* I. K. Vlachos	WiFi access points	Data communication over power lines to users' homes. Wireless-broadband over power lines (W-BPL) use medium voltage power lines as backhaul and wireless antennas for user access. Proposed system shows performance of raw 200 Mb/s after management data requirements such as smart grid applications.
MU-MIMO	Cellular	I. Latif N. L. Ratnayake	Sectorial Sectorial	Presents testbed LTE measurements with 800 MHz using multiple antennas to improve throughput and minimize interference. TV analog spectrum in Australia for sparsely populated areas leveraging spatial multiplexing gain.

*Due to the restriction on the number of references in this article, not all the works mentioned here are cited, and citations to these works and more details can be found in our technical report [5].

Table 2. Summary of deployment trends with case studies and on last-mile telecommunication in rural areas*.

with a Korean electricity company with data rates of 2 Mb/s. In the access segment, PLC uses LV distribution lines to provide access to houses or offices, connecting backhaul to a customer and in-building home wiring network to distribute the signal. Another field test provided throughput of 45 Mb/s (27 downstream and 18 upstream) over a distance of 600 m using repeaters.

Sometimes, an optical fiber network connecting two cities can be used to provide connectivity to the villages between them with the help of multiplexers and gateways. Methods to efficiently tap bandwidth from existing networks without jeopardizing the primary connections to connect rural areas need to be explored.

A hybrid architecture using MV lines and wireless (to replace LV in the access) is proposed where the MV lines act as backhaul links, and the last link of the network distribution uses WiFi/WiMAX wireless APs. A scheme of using such hybrid PLC for offering broadband access along a 107 km MV power grid in Larissa, a rural area in Greece, is described [5]. This technology is called hybrid wireless-broadband over power lines (W-BPL), and is based on the ubiquitous power grid and WiFi technology, with high potential for scalability and installation with data rates of 75–100 Mb/s. This study describes different quality of service (QoS) priority levels for applications such as smart grid, remote device management, and different network applications.

Multiple-Input Multiple-Output Wireless Networks

Spectral efficiency of standard technologies such as wireless local area network (WLAN), wireless local loop (WLL), and WRAN is quite limited. To achieve better spectral efficiency, a broad frequency spectrum or multi-user multiple-input multiple-output (MU-MIMO) technology is required. For APs equipped with multiple antennas, spectral efficiency and hence the capacity improves linearly as a function of the number of antennas without increasing total transmission power (spatial multiplexing gain).

Providing wireless broadband connectivity to Australia's rural areas is challenging due to a scattered population (2.7 persons/km^2). A novel approach is proposed to use analog TV spectrum with MIMO to leverage the spatial multiplexing gain (reviewed in [5]). These systems incur low interference and increased range due to beam forcing (a signal processing technique to reduce signal attenuation with distance) gain at the base station. In this work, the channel is modeled for varying weather conditions so that the transmitter array can adjust its properties to improve the spectral efficiency and the quality of the transmitted signal. This work also showcases the channel deployment steps.

Use of Alternate Infrastructure

In many countries, there is a substantial amount of alternative telecommunication infrastructures. For example, in India, Indian Rail (RailTel), the Gas Association of India Ltd. (GAILTel), and the national electricity distribution network (PowerGrid) have their own telecommunication networks. These alternative telecommunications networks (ATNs) operate their own in-house telecommunications systems, which have substantial built-in available capacity. Thus, while deploying a last-mile connectivity in rural areas, some of these ATNs' infrastructure can potentially be used in coalition with low-cost wireless solutions. There have been deployments using alternative networks in India [5].

Other Methods

In addition to these deployment trends, there are other technologies suitable for last-mile rural network connectivity such as free space aptics (FSO) and Zigbee (IEEE 802.15.4). More details on deployment trends based on these technologies are described in our technical report [5].

Discussion

Provisioning last-mile connectivity in rural areas has many practical challenges that need attention from the R&D community, as summarized below.

Smartphones as Computing and Network Devices

The fast market adoption of cell phones needs to be leveraged to provide appropriate computing and telecommunication services in rural areas. Like cell phones, smartphone technology will penetrate more in rural areas soon. A smartphone with data connectivity and basic features can deliver useful services, including a suitable interface for cloud computing and field-data collection and transfer. Smartphone applications suitable for rural users need to be developed to improve their knowledge, health, and business opportunities. R&D is required to provide blanket network coverage in rural areas to utilize smartphones and their applications.

Leveraging Existing Infrastructure and Technologies

Any existing infrastructure needs to be explored to provide telecommunication in rural areas from solutions utilizing telephone/electricity poles to mount WiFi/WiMAX or other antennas to act as relays/APs, to using solutions such as PLC. Sometimes, an optical fiber network connecting two cities can be used to provide connectivity between villages with the help of multiplexers and gateways. Methods to efficiently tap bandwidth from existing networks without jeopardizing the primary connections to connect rural areas need to be explored.

Also, while rural areas face problems that are significantly different from those of more developed areas, one can also exploit the benefits of the decreasing costs of standards-based mass-produced technology; for example, proliferation of inexpensive Android smartphones can help connect to the Internet. Some underserved rural areas could be served by hybrid methods that combine commercial and non-traditional approaches.

Reliable Connectivity

Just providing infrastructure for end-to-end connectivity in rural areas is insufficient. Mechanisms are also needed to make the network connection reliable and fault-resilient, especially for delay-intolerant services such as e-learning and emergency services.

Exploiting Emerging Trends: Open Source and Software-Defined Control

The emerging "open source" trend in telecom is a threat to many existing service providers and equipment vendors as it can make the deployment and operation of any network less expensive. Consider the example where multiple base stations are connected through the local wireless network, and calls are routed via private branch exchange (PBX) servers implemented in an open source framework, called Asterisk (www.asterisk.org). This system allows free calls within the local network and standard connections to outside callers using the satellite link. VillageCell in Africa uses free, open source solutions and off-the-shelf hardware, and hence the total deployment cost is minimal and the solution is scalable. Similar to the deployment discussed above, the voice and data from users are forwarded to the open source Asterisk PBX system, and the users use their existing cell phones for communication [15].

Also, many constraints that community-based networks impose on their design and operation can lead to novel research challenges and create new opportunities for service providers and equipment vendors. For example, the emerging trends of software-defined networking (SDN) and network functions virtualization (NFV) can be exploited to reduce the human intervention in these networks. The topic of autonomic networking during the last decade may now be more applicable here.

Concluding Remarks

The economic, social, and political life in the 21st century is increasingly becoming digital. It is important to provide computing and network services to underserved/remote communities to create sustainable growth and provide improved quality of life. We present how computing technologies can be used for various applications for rural empowerment. We identify several factors that significantly affect the design and implementation of last-mile telecommunication networks for rural areas. Deployment trends and case studies from different parts of the world show that "one size does not fit all." We outline possible R&D topics to develop innovative technology and sustainable business models that can improve computing and telecommunication capacity in rural areas.

Some important take-away messages follow. Choice of technology is crucial in setting up a network solution, and any existing infrastructure needs to be leveraged. Network solutions should provide reliable connectivity to grow a healthy and sustainable customer base. User density analysis needs to be conducted to identify locations of potential customers to motivate service providers to provide network services. Network solutions and business models need to be autonomous — operationally and financially self-sufficient — for rapid penetration in rural areas, and long-term sustenance. Besides telecommunication solutions, compelling applications suitable for rural communities need to be developed that can exploit the compute and data resources of today's cloud computing and networking paradigms such as SDN to motivate users to adopt and utilize available services.

> Besides telecommunication solutions, compelling applications suitable for rural communities need to be developed that can exploit the compute and data resources of today's cloud computing and networking paradigms such as software-defined networking (SDN) to motivate users to adopt and utilize available services.

References

[1] P. Goswami, P. Mahapatra, and Sw. Divyasukananda, "Bridging the Digital Gap in Rural India VIVEKDISHA: A Novel Experience," *Proc. Nat'l. Conf. Commun.*, 2013.

[2] A. Davis, A. Tall, and D. Guntuku, "Reaching the Last Mile: Best Practices in Leveraging the Power of ICTs to Communicate Climate Services to Farmers at Scale," *CGIAR Research Program on Climate Change, Agriculture and Food Security*, working paper no. 70, 2014, pp. 1–35.

[3] S. Surana *et al.*, "Deploying a Rural Wireless Telemedicine System: Experiences in Sustainability," *IEEE Computer*, vol. 41, no. 6, June 2008, pp. 48–56.

[4] A. Nandi, *Global Adoption of Mobile Phones – Trends and Interfering Factors*, M.S. thesis, Comp. Sci., UC Davis, Mar. 2014.

[5] S. Thota *et al.*, "Computing for Rural Empowerment: Enabled by Last-Mile Telecommunications (Extended Version)," tech. rep., UC Davis, http://networks.cs.ucdavis.edu/LastMile/LastMile_TechReport.pdf.

[6] R. Flickenger *et al.*, "Very Long Distance Wi-Fi Networks," *Proc., 2nd ACM SIGCOMM Wksp. Networked Systems for Developing Regions*, 2008.

[7] T. B. Chiyangwa and P. M. Alexander, "Rapidly Co-Evolving Technology Adoption and Diffusion Models," *Elsevier Telematics & Informatics*, vol. 33, no. 1, Feb. 2016 , pp. 56–76.

[8] D. Soman *et al.*, "Beyond Great Ideas: A Framework for Scaling Local Innovations," *Harward Business Rev.*, http://hbr.org/product/beyond-great-ideas-a-framework-for-scaling-local-i/an/ROT180-PDF-ENG, Sept. 2012.

[9] K. Ab-Hamid, C. E. Tan, and S. P. Lau, "Self-Sustainable Energy Efficient Long Range WiFi Network for Rural Communities," *Proc. IEEE GLOBECOM*, 2011, pp. 1050–55.

[10] J. Fitzpatrick, "Voice Call Capacity Analysis of Long Range WiFi as a Femto Backhaul Solution," *Computer Networks*, vol. 56, no. 5, Mar. 2012, pp. 1538–53.

[11] S. Sarkar, S. Dixit, and B. Mukherjee, "Hybrid Wireless-Optical Broadband-Access Network (WOBAN): A Review of Relevant Challenges," *IEEE/OSA J. Lightwave Tech.*, vol. 25, no. 11, Nov. 2007, pp. 3329–40.

[12] C. McGuire, M. R. Brew, F. Darbari, S. Weiss, and R. W. Stewart, "Enabling Rural Broadband via TV White Space," *Proc. 5th Int'l. Symp. Commun. Control and Signal Processing*, 2012.

[13] A. M. M. Altrad, W. R. S. Osman, and K. Nisar, "Modelling of Remote Area Broadband Technology over Low Voltage Power Line Channel," *Int'l. J. Computer Networks and Commun.*, vol. 4, no. 5, Sept. 2012.

[14] Z. A. Rajput *et al.*, "Evaluation of an Android-Based mHealth System for Population Surveillance in Developing Countries," *J. Amer. Med. Informatics Assn.*, vol. 19, no. 4, Feb. 2012, pp. 655–59.

[15] A. Anand *et al.*, "Villagecell: Cost Effective Cellular Connectivity in Rural Areas," *Proc. 5th Int'l. Conf. Info. and Commun. Technologies and Development*, 2012, pp. 180–89.

Biographies

Somen Nandi (snandi@ucdavis.edu) is the managing director and co-founder of Global HealthShare initiative (GHS) at the University of California, Davis. Before GHS, he spent 12 years in the biotechnology industry as a senior scientist and director, and developed several products that are currently in the market. He enjoys working with multidisciplinary and collaborative projects that enable him to appreciate the importance of multifaceted programs in diverse settings. He received his Ph.D. from the University of Calcutta (Bose Institute), India.

Saigopal Thota (sthota@ucdavis.edu) currently works as a software engineer at Cablevision Corporation. His research interests include last-mile networks, cloud-based service architectures for heterogeneous multi-device collaboration, and hybrid optical access networks. He received his Ph.D. in computer science from the University of California, Davis, in 2014, and his B.Tech. from Dhirubhai Ambani Institute of Information and Communication Technology, India, in 2009.

Avishek Nag (anag@ucdavis.edu) is a research associate at CONNECT, Ireland's National Research Centre for Future Networks and Communications. His research interests include cross-layer optimization in wired and wireless networks, network reliability, mathematics of networks, network virtualization, and software-defined networks. He received his B.E. (honors) from Jadavpur University, Kolkata, India, in 2005, his M.Tech. from the Indian Institute of Technology (IIT), Kharagpur, in 2007, and his Ph.D. from the University of California, Davis, in 2012.

Sw. Divyasukhananda (rkmsadhu@gmail.com) is the coordinator, VIVEKDISHA, an ICT-based network of Ramakrishna Mission Vivekananda University, India. His mission is to spread quality education at the high school level and beyond, and to break the digital divide between urban and rural areas by conducting online classes and providing telemedicine services.

Partha Goswami (partha@cc.iitkgp.ernet.in) is a senior networking engineer at IIT Kharagpur. He has worked for 21 years in the area of computer networking. He received his B.Sc. in physics (Honors) and B.Tech. in radio physics and Electronics from the University of Calcutta in 1991 and 1994, respectively. He completed his M.Tech. (2006) from the School of Information Technology and Ph.D. (2016) from the Department of Electronics and Electrical Communication Engineering at IIT, Kharagpur.

Ashwin Aravindakshan (aaravind@ucdavis.edu) is an associate professor of marketing at the University of California, Davis. His research focuses on the analytics of consumer behavior and the optimization of marketing actions. His work has appeared in *Operations Research*, *Management Science*, the *Journal of Marketing Research*, and *Marketing Science*, among others. He received a B.Tech. in aerospace engineering from IIT, Madras, and a Ph.D. in marketing from the University of Maryland, College Park.

Raymond L. Rodriguez (rlrodriguez@ucdavis.edu) is a professor of molecular and cellular biology and Executive Director of the GHS initiative at the University of California, Davis. His research interests include complex biological systems from environment x genome interactions to global health and disease.

Biswanath Mukherjee [F] (bmukherjee@ucdavis.edu) is Distinguished Professor of Computer Science at the University of California, Davis. He received his B.Tech. degree from IIT, Kharagpur (1980) and his Ph.D. from the University of Washington, Seattle (1987). He was General Co-Chair of IEEE/OSA OFC '11, Technical Program Co-Chair of OFC '2009, and Technical Program Chair of IEEE INFOCOM '96. He is Editor of Springer's Optical Networks Book Series.

Consumer Communications and Networking

Supporting Consumer Services in a Deterministic Industrial Internet Core Network

Ted H. Szymanski

A convergence is occurring in the networking world. Industrial networks currently provide deterministic services in robotic factories and aircraft, while the best effort Internet of Things provides best effort services for consumers. The author argues that a convergence should occur, and that a future converged industrial Internet of Things (IIoT) should support both best effort and deterministic services, with very low latency and jitter.

Abstract

A convergence is occurring in the networking world. Industrial networks currently provide deterministic services in robotic factories and aircraft, while the best effort Internet of Things provides best effort services for consumers. We argue that a convergence should occur, and that a future converged Industrial Internet of Things (IIoT) should support both best effort and deterministic services, with very low latency and jitter. This article presents the design of a deterministic IIoT core network consisting of many simple deterministic packet switches configured by an SDN control plane. The use of deterministic communications can reduce router buffer sizes by a factor of ≥ 1000, and can reduce end-to-end latencies to the speed of light in fiber. A speed-of-light deterministic core network can have a profound impact on virtually all consumer services such as multimedia distribution, e-Commerce, and cloud computing or gaming systems. Highly aggregated video streams can be delivered over a deterministic virtual network with very high link utilization (≤ 100 percent), very low packet jitter (≤ 10 μs), and zero congestion. In addition to improving consumer services, a converged deterministic IIoT core network can save billions of dollars per year as a result of significantly improved network utilization and energy efficiency.

Introduction

The existing best effort Internet of Things (BE-IoT) suffers from congestion and provides inefficient best effort service for consumers. It provides no guarantees for the bandwidth, delay, or jitter of a consumer's Internet connection(s), and it is typically overprovisioned to operate at light loads, to reduce delay, jitter, and packet loss rate. This over-provisioning costs service providers several billions of dollars per year in excess capital costs and energy costs, and large delays still occur frequently during times of congestion. As a result of congestion, the BE-IoT cannot support the demanding machine-to-machine (M2M) communications required in robotic factories, airplanes, and space craft. We argue that the future converged industrial Internet of Things (IIoT) should support both best effort services for consumers and deterministic services for M2M communications, where the end-to-end delay, jitter, and packet loss rate can be deterministically bounded.

General Electric (GE) coined the term Industrial Internet to acknowledge the growing importance of connecting industrial machines rather than humans. Industrial automation will use the IIoT to enable a new wave of robotic manufacturing, by interconnecting industrial sensors, control systems, and robots. GE envisions that the transformation to industrial automation may impact the world on the same scale as the industrial revolution of the 19th century. It estimates that industrial automation may increase worldwide GDP by $15 trillion by 2030 by reducing costs and waste, and improving manufacturing processes. GE also estimates that the IIoT may control about $82 trillion of industrial GDP by 2030, representing about half of the world's GDP. In March 2014, five companies (GE, Cisco, AT&T, IBM, and Intel) formed the Industrial Internet Consortium to advance the technologies, and in June 2015 the Consortium included 160 companies, indicating strong industrial support.

Reduction of the large Internet latencies has received significant attention lately. In 2013, the Association of Computer Manufacturers (ACM) and the Internet Society held a workshop on reducing Internet latencies, which concluded that unnecessary delays should be removed from every layer of the protocol stack [1]. A recent ACM paper, "The Internet at the Speed of Light," shows that Internet latencies are typically 10× to 100× larger than the minimum delays due to the speed of light in fiber [2]. They argue that a speed-of-light Internet would be a "technological leap" forward that could fundamentally transform computing. For example, a speed-of-light Internet could transform cloud computing or gaming systems, both multi-billion-dollar industries, by dramatically increasing the size of the reachable population (of machines or people) given a fixed latency. In 2014, the International Telecommunication Union (ITU) began to explore the impact of a Tactile Internet network, with a goal to reduce end-to-end latencies to 1 ms. They argue that the Tactile Internet would add a new dimension to human-machine interaction and revolutionize M2M interaction [3].

Large Internet latencies lead to very high costs in the e-Commerce industry. In 2014, global e-Commerce revenue was about US$1.2 trillion.

The author is with McMaster University.

0163-6804/16/$25.00 © 2016 IEEE

A 100 ms latency penalty can reduce sales for Amazon by 1 percent, and similar figures have been reported for Bing and Google [2]. Amazon's sales revenues were US$89 billion in 2014, and a 1 percent loss represents over US$1 billion in 2015. According to Akamai, a quick page load time is a key factor in a consumer's loyalty to an e-Commerce site, as 40 percent will wait no longer than 3 s before abandoning the site.

Large Internet latencies also lead to very high costs in the financial services industry. A 1 ms increase in latency can reduce revenue by US$100 million per year for firms performing high frequency automated stock trading. Internet latencies can be reduced by deploying new fiber, but the cost is prohibitive. For example, the cost of deploying new fiber under the Arctic Circle to reduce the London-to-Tokyo latency by 60 ms is US$1.5 billion.

According to Sandvine Networks, large-scale video distribution from services such as YouTube and NetFlix currently consumes about 50 percent of the continental U.S. core bandwidth at peak times. This figure is expected to rise to potentially 90 percent in the future.

We believe that a convergence of the best effort and deterministic communications paradigms into a single unified network should occur. This article first presents the design of a deterministic IIoT core network based on a network of simple deterministic packet switches controlled by a software defined networking (SDN) control plane [4]. The packet switches can operate at layer 2 or 3, as shown in Fig. 1. Our SDN control plane can program thousands of deterministic virtual networks (VNs) into the IIoT core to distribute highly aggregated video streams, as shown in Fig. 2. Our deterministic IIoT design has three unique features:

- It can provably operate all Internet links at 100 percent loads.
- It can simultaneously reduce end-to-end transport delays to the speed of light in fiber.
- The complexity of scheduling traffic through the switches with low jitter is not NP-Hard [4].

We show that the ability to operate the future deterministic core network at 100 percent capacity can lead to potential capital cost savings of US$37 billion per year.

This article is organized as follows. We discuss the evolution to a converged IIoT network. We present the design of a deterministic packet switch for layer 2 or 3. We present a deterministic U.S. core network and its performance. We explore the distribution of aggregated video over the converged IIoT. We conclude the article.[1]

The Evolution to Deterministic Services

The existing BE-IoT poses several challenges for industrial automation and the consumer services industry. The BE-IoT suffers from congestion, which causes:

- Excessively high end-to-end delays potentially as large as 50–500 ms
- Potentially high packet loss rates of 5–50 percent, unless the network is significantly overprovisioned [4]

The Internet Engineering Task Force (IETF) acknowledges that overprovisioning lowers the utilization of the BE-IoT infrastructure to typically below 50 percent [5, 6]. The IETF has ruled out overprovisioning as a means to achieve deterministic services, and aims to achieve at least 50 percent link utilizations for deterministic traffic flows in the future Internet.

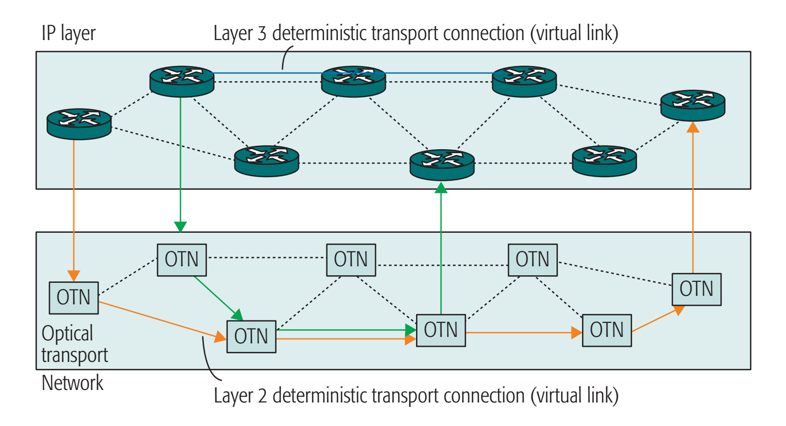

IP layer
Layer 3 deterministic transport connection (virtual link)
OTN
OTN
OTN
OTN
OTN
OTN
OTN
OTN
Optical transport Network
Layer 2 deterministic transport connection (virtual link)

Figure 1. A layer 3 network of IP routers, with a layer 2 optical transport network (OTN) underlay. "Deterministic transport connections" (virtual links) can be embedded into each layer.

Current BE-IoT routers typically use a bandwidth-delay product buffer sizing rule, which provides buffers for about 250 ms of data per IO port to provide congestion control for worst case scenarios [4, 7]. A router with 400 Gb/s links has buffers for about 100 Gbits of data per IO port (in the worst case), equivalent to about 8.3 million maximum-size IP packets. Referring to Fig. 2 and assuming 400 Gb/s links, the BE-IoT router in Chicago will have buffers for about 32 million IP packets. These large buffers increase BE-IoT router complexity, costs, power consumption, and failure rates, and play a key role in the BE-IoTs excessive delays during times of congestion.

ATM and MPLS-TE Core Networks

A deterministic traffic flow is immune to congestion and interference from all other traffic flows, and can also be called a guaranteed rate (GR) or constant bit rate (CBR) flow. In the 1990s the international community developed the asynchronous transfer mode (ATM) standard with CBR service (in principle). Unfortunately, the problem of scheduling CBR traffic flows through an input-queued packet switch with minimum delay and jitter is a well-known NP-Hard problem; see [4, 8–10]. The ATM standard did not solve the NP-Hard switch scheduling problem and could not provide a true deterministic service [4]

The ATM standard was eventually abandoned. Multiprotocol label switching with traffic engineering (MPLS-TE) was developed shortly thereafter to offer improved service. However, the MPLS-TE standard also did not solve the NP-Hard switch scheduling problem and could not provide a true deterministic service [4].

MPLS-TE exists today, but it does not provide a true deterministic service for industrial automation, robotic manufacturing, and mission-critical M2M communications [4].

[1] A related video, "A Speed of Light Deterministic Industrial Internet of Things," can be found at https://youtu.be/cXA0HEjKRPY

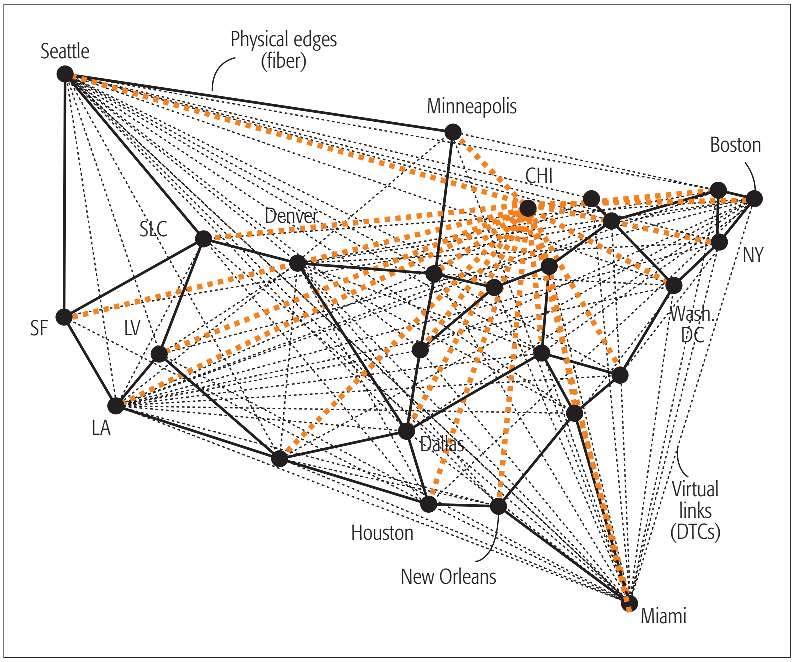


Figure 2. A deterministic U.S. IIoT core network with virtual links originating at six cities (Seattle, Los Angeles, Denver, Chicago, Boston, and Miami). A virtual network supporting video distribution from Chicago is highlighted (in red).

Existing Industrial Networks

Proprietary industrial networks offering low-latency deterministic M2M services have existed for years in industrial automation and the avionics industry. However, the proprietary nature of these networks has increased costs and limited wide-scale deployment. For example, Airbus developed the patented Avionics Full-Duplex Switched Ethernet (AFDX) network for the Airbus 380. The A380 requires over 500,000 m of control wires. Unfortunately, in 2006 wiring problems (the wires were a few inches too short) delayed the A380 project, leading to cost overruns of €2 billion. The proprietary nature of the control wires meant that low-cost replacements were not readily available. In 2014, wiring problems delayed the new U.S. Air Force KC-46 refueling tanker project leading to cost overruns of US$1.5 billion. The IEEE recently developed the deterministic Ethernet standard to provide an open low-cost standard to support both best effort and deterministic M2M services, to avoid similar problems recurring.

The Deterministic Ethernet Access Network

To address the need for deterministic M2M services in factories, vehicles, trains, planes, and audio/video applications, the IEEE developed the 802.1Q standard for Deterministic Ethernet to provide both deterministic and best effort services on a single Ethernet link [11]. The standard requires that a packet must be delivered within a deterministic time bound, but for flexibility it does not specify any scheduling algorithms. Typically, an application will explicitly reserve times for data transmissions on the Ethernet broadcast medium to achieve a deterministic delay bound. The IEEE standard requires that all applications on the broadcast medium are synchronized, potentially to within nanoseconds or microseconds of accuracy, to avoid packet collisions. The IEEE also added 3 bytes to the basic Ethernet packet size to allow for the identification of 16 million virtual networks.

Layer 2 networks are usually small, and are typically interconnected with service provider bridges and backbone bridges. Traditionally, a layer 3 IP core network interconnects "islands" of smaller layer 2 networks. The IEEE is currently looking at the requirements for providing deterministic services in larger layer 2 networks, such as bridged and switched Ethernet networks and rings, to support real-time M2M services. However, the introduction of switches significantly complicates the provisioning of deterministic services, since the problem of scheduling deterministic traffic flows through one switch with minimum delay and jitter is NP-Hard in general [10]. Our scheduling algorithms can also be used to program deterministic layer 2 bridges.

In Fig. 1, our layer 2 OTN can span a continent, where each simple deterministic packet switch can ideally fit on a field programmable gate array (FPGA). The layer 2 switches must obey strict deterministic packet forwarding schedules and could use any transport-oriented packet format, for example, the Deterministic Ethernet or carrier Ethernet packet formats. Hence, the network in Fig. 1 can be viewed as IP-over-Carrier-Ethernet-over-dense wavelength-division multiplexing (DWDM).

The Wireless TSCH Access Network

The IETF has created a Working Group, 6TiSCH, to incorporate IEEE's time synchronized channel hopping (TSCH) wireless standard into the IP infrastructure [12]. The standard will provide deterministic real-time M2M services for "last-mile" wireless access networks supporting IPv6. The TSCH standard allows a wireless node to explicitly reserve time slots for transmission on several frequency-based channels. The transmissions of a traffic flow will typically experience extensive frequency hopping to mitigate the effects of wireless fading and interference in any one frequency. However, for flexibility the standard does not specify the scheduling algorithms to be used.

IETF Activities in Deterministic Networks

In October 2015, the IETF approved the Deterministic Networking Group to explore the feasibility of adding deterministic services to the BE Internet network (as a work in progress). The IETF has published a draft Deterministic Networking Problem Statement [5] and a Deterministic Forwarding Per Hop Behavior (PHB) [6] draft for use with the differentiated services (DiffServ) service model. These drafts specify an abstract model rather than a detailed technical solution. The drafts require that the packets in a deterministic flow must receive deterministic service in each Internet router, but for flexibility they do not specify the router architecture or any routing/scheduling algorithms. The IETF drafts require that all routers are synchronized, to within 10 ns–10 μs of accuracy. This tight synchronization is a potential problem for a deterministic network that spans a continent as shown in Fig. 2 (our approach solves this problem).

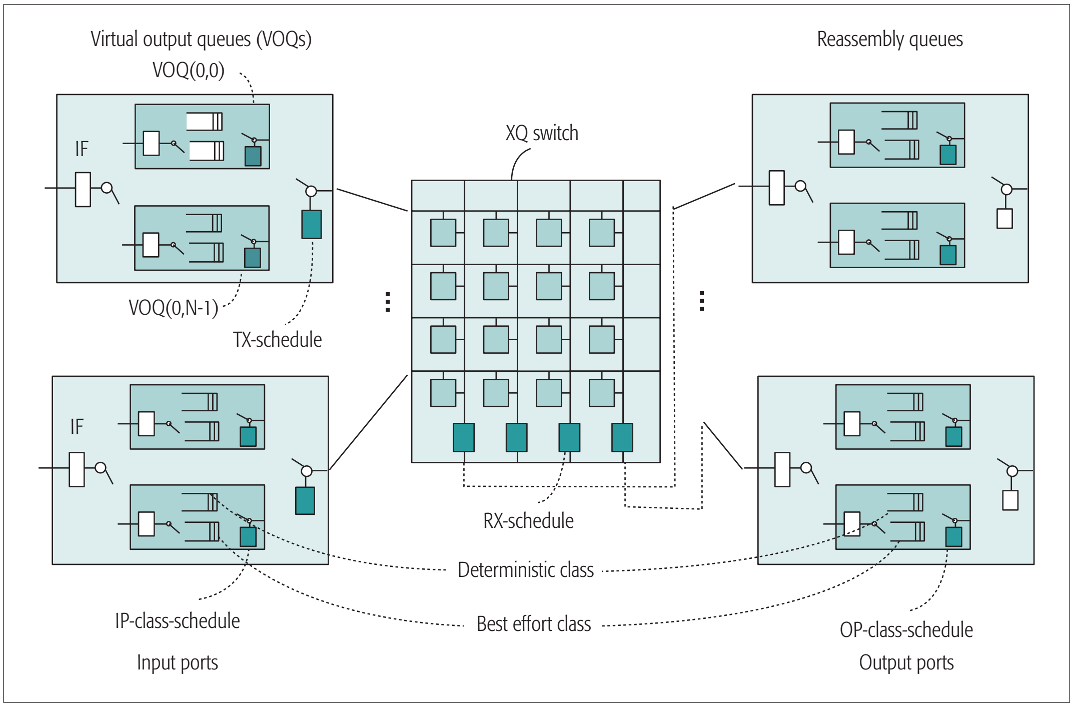


Figure 3. Basic deterministic switch with combined input, crosspoint and output queueing (CIXOQ).

In late 2015, the Industrial Internet Consortium published a draft Industrial Internet Reference Architecture. The first draft identifies the most important architectural issues and it is "broad rather than deep." The architecture does not mention deterministic communications or time-synchronization requirements, but it does discuss the use of prioritization to achieve better service for M2M flows.

The IETF has proposed several use cases for deterministic communications, including:

- Professional audio over the Internet
- Deterministic radio access networks
- Deterministic mobile networks
- Deterministic control for utilities such as the smart power grid

The existing power grid distributes vast amounts of power over a network of high-voltage transmission lines. The ability to increase transmission line utilizations by 10 percent can lead to potential capital cost savings of several billion dollars [13]. However, the future smart power grid will require a very fast control system. According to the IETF, jitter of less than 250 μs and end-to-end delays of less than 4–10 ms are needed [13]. The deterministic U.S. IIoT network shown in Fig. 2 can meet a 10 ms delay constraint over distances of about 2000 km, and the jitter is less than 10 μs.

The Industrial Internet and Tactile Internet Projects

In late 2015, the Industrial Internet Consortium published a draft Industrial Internet Reference Architecture. The first draft identifies the most important architectural issues and is broad rather than deep. The architecture does not mention deterministic communications or time synchronization requirements, but it does discuss the use of prioritization to achieve better service for M2M flows. In this article, we argue that deterministic communications offers several benefits over best effort communications using prioritization, and present a deterministic Industrial Internet core network.

In 2014, the ITU began a project on the Tactile Internet to describe a future Internet network with exceptionally low end-to-end latency, and high availability, reliability and security, for applications including industrial automation and smart transportation systems. This project does not mention deterministic communications or time synchronization requirements. The Tactile Internet project has the same goals as the Industrial Internet project.

A Deterministic Packet Switch

Packet switches use several types of queueing, including input queueing (IQ), output queueing (OQ), and combined input and output queueing (CIOQ). An $N \times N$ OQ switch can achieve 100 percent throughput with minimum delay; however, it requires an internal speedup of N to remove all contention for output ports, which increases costs and power use. Large OQ switches are intractable and are rarely used [4].

An IQ or CIOQ switch can achieve 100 percent throughput with no internal speedup, or with a small Internet speedup of typically 2 or 4. However, complex scheduling algorithms are needed to schedule the packets through the switch with 100 percent throughput and without contention [8, 9]. The problem of scheduling deterministic traffic flows through an IQ or CIOQ switch with no speedup, 100 percent throughput, and minimum delay and jitter is NP-Hard; see [4, 10].

Figure 3 illustrates a simple packet switch that supports deterministic traffics flows with 100 percent throughput, and deterministic delay and jitter guarantees [14]. The switch adds crosspoint queues (XQs) to the basic CIOQ switch, yielding a combined input, crosspoint, and output queueing (CIXOQ) switch. The majority of buffering occurs at the input ports (IPs) and output ports (OPs); the XQs are very small, and exist only to simplify the scheduling algorithms. Variable-size Internet packets arrive at the IPs, and are typically fragmented into fixed sized cells (with 64 or 128 bytes) for transmission through the switch. The variable-size Internet packets are reassem-

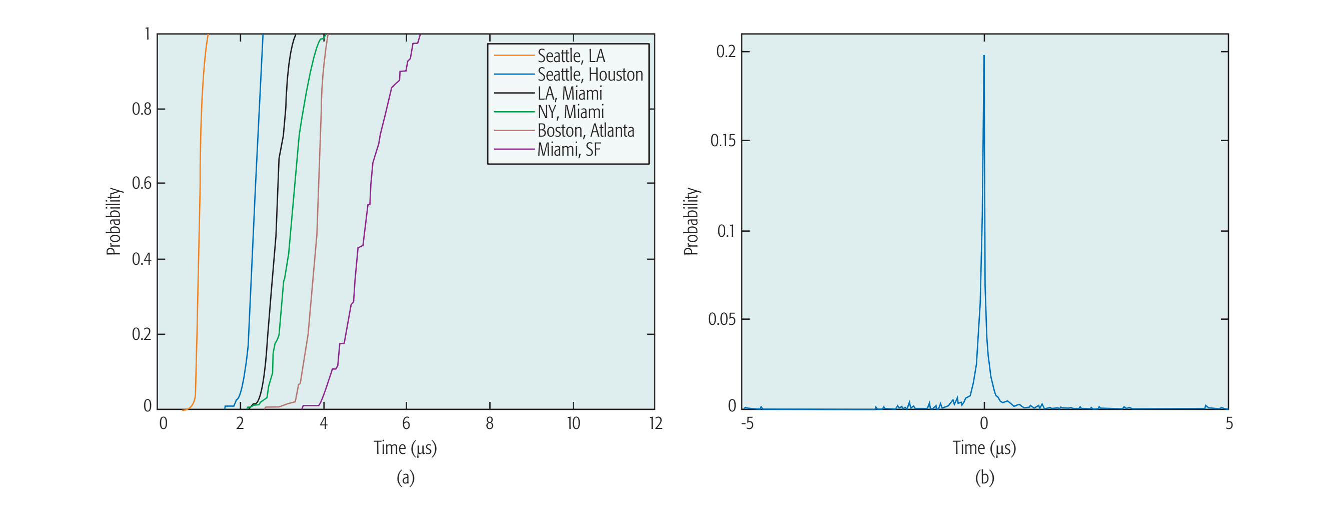


Figure 4. a) End-to-end queueing delay CDF for selected flows in the USA backbone; b) jitter distribution for all flows in the U.S. backbone.

bled at the output side of the switch. In an $N \times N$ switch, each IP has N virtual output queues (VOQs), where VOQ(i,j) stores data which arrives at IP(i) and is going to OP(j). Each VOQ in Fig. 3 supports two prioritized traffic classes, the deterministic and best effort classes.

However, a VOQ can be partitioned to support many prioritized traffic classes, including the three existing DiffServ traffic classes, expedited forwarding (EF), assured forwarding (AF), and DE), and a new deterministic class. Another new traffic class can also be created to handle short TCP/IP control packets (i.e., TCP acknowledgment [ACK] packets and socket open/close connection packets) with expedited guaranteed rate (GR) service.

Let each $N \times N$ CIOQ or CIXOQ switch have an $N \times N$ matrix of guaranteed traffic rates to be supported between the input and output ports. Reference [4] presents a very fast recursive scheduling algorithm, which can schedule the transmission of packets through a CIOQ switch with near-minimal delay and jitter. The algorithm mathematically recursively decomposes the $N \times N$ matrix of guaranteed traffic rates to achieve a very low-jitter transmission schedule with 100 percent throughput.

When the XQs are added to the CIOQ switch, as shown in Fig. 3, the scheduling algorithm is simplified. Each row of the $N \times N$ traffic matrix can be processed in isolation to compute a TX-Schedule for each IP. (Each row of the $N \times N$ matrix is a $1 \times N$ vector, which can be processed using the recursive scheduling algorithms in [4, 14].) At each IP, the TX-Schedule identifies a VOQ to be serviced for each time slot of a scheduling frame. The TX-Schedule provides each IP with a guaranteed rate of transmission, from the VOQs into the XQs of the switch. Each column of the $N \times N$ traffic matrix can also be processed to compute an RX-Schedule for each OP. The RX-Schedule specifies the XQ to be serviced in each column of the XQ switch for each time slot of a scheduling frame. The RX-Schedule provides each OP with a guaranteed rate of reception from the XQs into the OQs of the switch. When an IP receives service in a time slot, an *Input-Class-Schedule* can specify the traffic class or traffic flow to be serviced. At the OPs, once packets are reassembled, an optional *Output-Class-Schedule* can specify the traffic class or traffic flow to be serviced.

The Deterministic Schedules

The switch in Fig. 3 can reserve time slots for the transmissions of every deterministic traffic flow in a scheduling frame with F time slots. A scheduling frame length of $F = 1024$ can allocate bandwidth in increments of 0.1 percent of the line rate. Using a 400 Gb/s line rate and $F = 1024$, each time slot reservation will reserve 400 Mb/s. Given a traffic rate matrix, the schedules can be computed in microseconds. The traffic demands for deterministic flows change relatively slowly, over seconds or minutes, and hence the schedules can be computed once and stored in lookup tables and reused until the traffic demands change. The shaded boxes in Fig. 3 represent lookup tables. The routers do not need to be synchronized to microseconds of accuracy, as these schedules can be circularly rotated by arbitrary amounts and still retain the deterministic delay and jitter bounds. This ability to circularly rotate schedules is very important, since all the routers or switches in the U.S. core network in Fig. 2 need not be synchronized.

A Deterministic U.S. IIoT

Figure 2 illustrates a deterministic U.S. IIoT core network, with 26 nodes (cities) and 86 edges. The bold lines represent optical fiber links between cities. The dotted lines represent congestion-free deterministic transport connections (DTCs) between cities.

Our SDN control plane can program many virtual networks (VNs) into layer 3, as shown in Fig. 2. A VN is composed of many virtual links (VLs), where each VL represents a DTC, which passes through many routers. A router views a DTC as a congestion-free one-hop VL between remote cities, as shown by the dotted lines in Fig. 2. Our SDN control plane can configure the deterministic connections in layer 3 by configuring each router with several deterministic forwarding schedules. Packets of a DTC will be forwarded along a fixed path of Internet routers using these deterministic forwarding schedules, resulting in near-minimal buffer sizes and queueing latencies. Our SDN control plane uses a Max-Flow Min-Cost routing algorithm [15], without relying on sub-optimal best effort IP routing algorithms. It can create single-path or multi-path DTCs, with redundancy for improved reliability.

Our SDN control plane can also embed many VNs and VLs into an optional layer 2 underlay network of simple packet switches called the optical transport network (OTN), as shown in the bottom part of Fig. 1. Each VL in layer 2 can bypass several IP routers in layer 3, which will significantly improve energy efficiency. Current IP routers consume between 5 and 10 nJoules per bit transmitted, while state-of-the-art layer 2 packet switches consume about 250 pJoules per bit, resulting in an energy savings of a factor of about 30. The use of deterministic connections in layers 2 and 3 can help meet the aggressive energy efficiency targets specified by the Greentouch consortium (www.greentouch.org).

In Fig. 2, our SDN control plane programmed 300 VLs into the IIoT. Six cities selected at random (Seattle, Los Angeles, Denver, Chicago, Boston, and Miami) each have 50 VLs, with 25 VLs going to/from the other cities. Each of these six cities can reach any other city over a one-hop VL. (In Fig. 2, it is straightforward to embed a fully connected network where every pair of cities is interconnected with two VLs.) An IP router can also use VLs in its Open Shortest Path First (OSPF) and Border Gateway Protocol (BGP) routing algorithms to support best effort traffic. These routing algorithms often minimize the number of hops, and they can be modified to use the VLs, which are viewed as one-hop logical connections between cities.

Experimental Results with 92 Percent Loads

In our tests, a scheduling frame with 1024 time slots was used. Each time slot was sufficient to transmit a maximum-size IP packet over an edge. Assuming 400 Gb/s edges and 1500-byte IP packets, a time slot consists of 30 ns. (A 400 Gb/s edge may consist of 4 parallel 100 Gb/s channels, in which case a time slot consists of 120 ns).

The IIoT network performance is deterministic, and was determined using three methods, which were all in agreement;

1. Reference [4] presents theoretical bounds on the end-to-end latencies and jitter.
2. A software simulator was developed to simulate the deterministic system.
3. An FPGA hardware testbed was developed where 26 simple routers were synthesized onto an Altera FPGA, and the performance was measured in hardware.

The hardware testbed can transmit packets at a rate exceeding 400 million packets/s. The hardware testbed and software simulator yield identical deterministic results.

Figure 4a illustrates the cumulative distribution function (CDF) of the end-to-end queueing delay between several cities, expressed in time slots. This figure does not include the fiber latency. The queueing delays in Fig. 4a are all less than 10 μs. Using standard single-mode fiber, the speed of light is about 200 km/ms. Consider the VLs between Los Angeles and Miami. The length of the fiber between these cities is at least 3800 km, depending on the physical path. The fiber latency is therefore 19 ms. The end-to-end queueing delay along the VL (≤ 10 μs) is over 1000 times smaller than the end-to-end fiber delay (≥ 19 ms).

Figure 4b illustrates the probability distribution of the jitter of the packets leaving a VL (averaged over all VLs in the U.S. network). The jitter is defined as the time difference of two consecutive departing packets in a given VL minus the ideal time between packets in the VL. According to Fig. 4b, most packets are delivered with a jitter ≤ 1 μs. According to theory, given a VL with a provisioned rate of 10 Gb/s and maximum-size IP packets, the maximum jitter is about 1.2 μs [4]. These jitter times, measured in microseconds, are exceptionally small when compared to the end-to-end fiber delays in the U.S. backbone network, measured in milliseconds.

According to our testbed the switch at Chicago buffers less than 50 packets, even at 93 percent average link loads. A BE-IoT router at Chicago with 400 Gb/s links would have a worst case buffer size of about 32 million packets [4, 7]. The use of deterministic packet switching, combined with our very low-jitter scheduling algorithm, has reduced the worst case buffer sizes by a factor exceeding 100,000 times.

Large-Scale Video Distribution

In this section, we explore large-scale video distribution over the future deterministic IIoT using simulations. Assume a Netflix data center in Chicago distributes video to several cities. Our SDN control plane can program a VN into the IIoT to support video distribution, with VLs from Chicago, to all other cities, as shown in Fig. 2. Netflix has about 35 million subscribers in the United States, and alone accounts for about 33 percent of the U.S. download bandwidth in peak hours. Each VL will typically carry between 1000 and 100,000 video streams, depending on the time of day. The provisioned rate of the VLs can be updated by an autonomic controller in the SDN control plane every 15 minutes (or as needed).

Video encoders typically use the standard three-level group of pictures (GOP) format. Each GOP consists of a large independent (I) frame, followed by several optional smaller predictive (P) frames, where several small bi-predictive (B) frames may exist between the P frames.

Figure 5a explores the aggregation of multiple low-bit-rate single-layer scalable video coding (SVC) video streams for mobile devices (i.e., tablets and smartphones). No buffering to smooth the traffic is assumed in Fig. 5. A single-layer SVC video stream called "Gandhi" is used, with a G16B15 GOP format. The G16B15 GOP format has one independent I frame followed by 15 smaller B frames. It has 53,968 frames, with a screen size of 352 × 288 pixels, a rate of 7.5 frames/s, with an average bit rate of 18.5 kb/s. To generate multiple video streams for aggregation, the same video stream was circularly rotated by a random amount before aggregation. Referring to the top row in Fig. 5a, the single video stream

Netflix has about 35 million subscribers in the USA, and alone accounts for about 33 percent of the USA download bandwidth in peak hours. Each VL will typically carry between 1,000 and 100,000 video streams, depending upon the time of day. The provisioned rate of the VLs can be updated by an autonomic controller in the SDN control-plane every 15 minutes (or as needed).

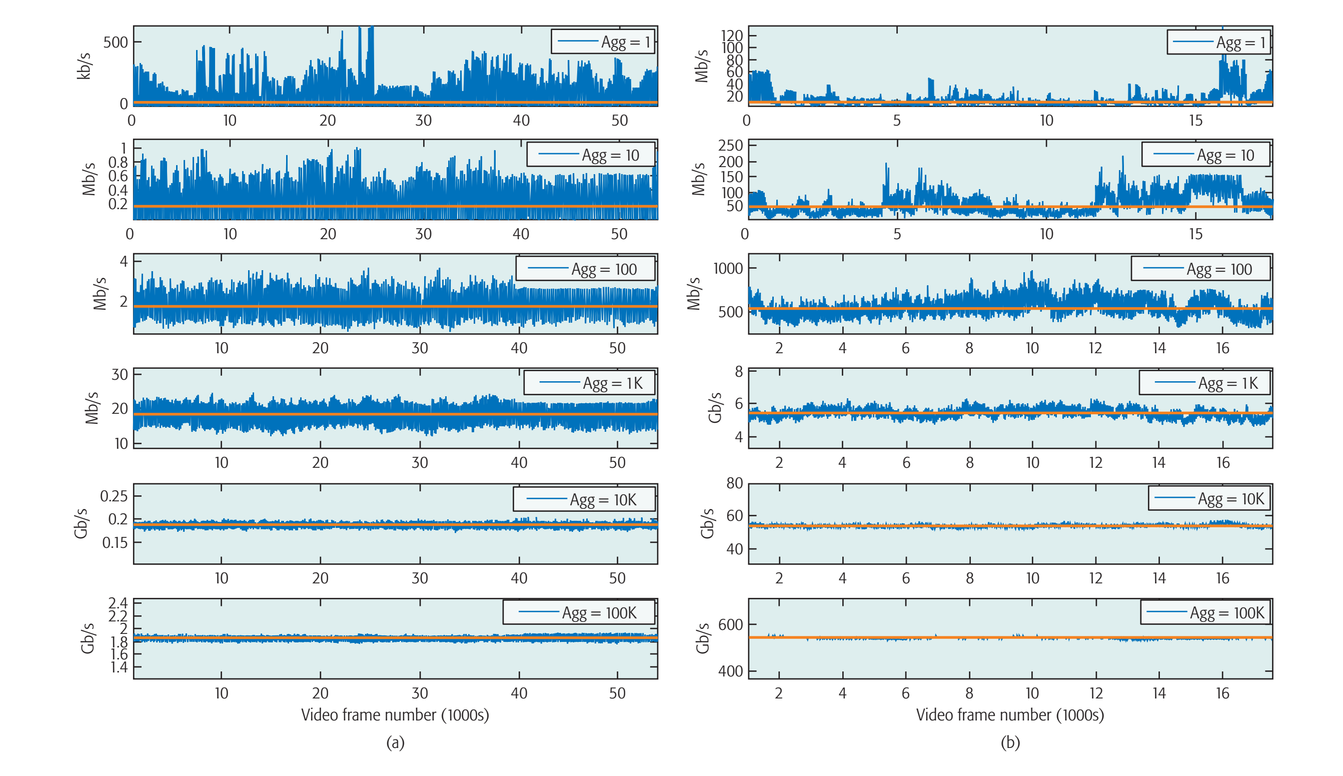


Figure 5. Instantaneous bandwidth vs. video frame number, for various degrees of aggregation: a) SVC video "Gandhi"; b) V9P 4K UHD video "Tears of Steel."

is quite bursty, with a mean bit rate of 18.5 kb/s and a peak bit rate of about 500 kb/s. When 10K streams are aggregated, the mean bit rate is 185 Mb/s and the peak bit rate is about 190 Mb/s, a considerable reduction in burstiness. We have aggregated several other SVC videos and observed the same behavior. A VL can transport thousands of SVC video streams with very high link utilizations, given the reduction in burstiness that occurs with aggregation.

Figure 5b explores the aggregation of multiple ultra high definition (UHD) 4K video streams for home TV. There are several encoders for UHD video, including the H.264, H.265, and VP9 encoders. A single video, the UHD VP9 4K "Tears of Steel" video, is used, with a GOP format of G24B0. It has 17,952 frames, with a screen size of 4096 × 1744 pixels, a rate of 24 frames/s, and an average bit rate of 5.414 Mb/s. To generate multiple video streams for aggregation, the same video stream is circularly rotated by a random amount. Referring to the top row of Fig. 5b, the single video stream is quite bursty, with a mean bit rate of 5.4 Mb/s and a peak bit rate of about 120 Mb/s. When 10K streams are aggregated, the mean bit rate is 54 Gb/s and the peak bit rate is about 56 Gb/s, a considerable reduction in burstiness. A VL can transport thousands of UHD video streams with very high link utilization, given the reduction in burstiness that occurs with aggregation.

The northbound traffic leaving a data center going to a remote city represents the aggregation of thousands of video streams. A token-bucket-based video shaper queue (VSQ) can be used at each data center to further smooth an aggregated stream before transmission over a VL. Our simulations indicate that an aggregated stream of 1000 videos (or more) can be delivered with queueing delays in the VSQ of ≤ 2–4 ms, with link utilizations of 95 percent [15]. In other words, significant overprovisioning is not needed. The southbound traffic arriving at a destination data center from the core network represents the aggregation of thousands of video streams. A video playback queue (VPQ) can be used to demultiplex the smoothed aggregated stream into multiple bursty video streams for distribution over a local area network. This playback queue will incur a similar small delay of typically 2–4 ms [15]. In the continental U.S. network shown in Fig. 2, the queueing delays in the VSQ and VPQ are much smaller than the fiber latencies.

The IETF has ruled out the use of overprovisioning to support deterministic services, and states that link utilizations of at least 50 percent should be supported for deterministic traffic in the future Internet [5, 6]. According to 2013 and 2014 annual reports, the annual sales of best effort hardware (routers, switches, wireless nodes) from Cisco, Huawei, Ericsson, and Alcatel-Lucent can be estimated at US$22, US$23, US$14.2, and US$15 billion, respectively, for a total of US$74 billion annually. Assuming a 50 percent link utilization, half of this annual hardware expenditure is effectively unused, and the unnecessary capital costs of underutilized networks can reach US$37 billion annually. Hence, it is desirable to achieve higher utilizations, well above 50 percent and approaching 100 percent, for deterministic traffic. This article demonstrates the technologies to achieve up to 100 percent utilization for deterministic traffic flows, using simple low-cost CIOQ or CIXOQ switches, which can lower the excess capital costs and energy costs of a future deterministic core network significantly.

Conclusion

The Internet network has used an inefficient best effort communications paradigm for the last 40 years, incurring excessive delays, capital costs, and energy costs. This article proposes a deterministic Industrial Internet of Things core network consisting of many simple deterministic packet switches controlled by an SDN control plane. Our SDN control plane can program thousands of deterministic virtual networks into the core network to provide each consumer service with its own dedicated congestion-free VN with exceptionally low latency and jitter. Highly aggregated video streams can be delivered over the continental United States with very low end-to-end latency determined by the speed of light in fiber, with jitters less than 10 μs, and with up to 100 percent link utilizations. By achieving 100 percent link utilizations rather than the 50 percent targeted by the IETF, the proposed deterministic network can save potentially US$37 billion in capital costs annually. A speed-of-light deterministic core network can have a profound impact on virtually all consumer services such as multimedia distribution, e-Commerce, and cloud computing or gaming. It can also pay for itself quickly, due to its significantly improved utilization and energy efficiency. The deterministic technologies proposed in this article can also provide deterministic services in metro area networks, data center networks, and supercomputer networks. We believe that a future converged deterministic Internet of Things that combines the best effort and deterministic communications paradigms can fundamentally transform computing and consumer services in the 21st century.

References

[1] M. Ford, “Workshop Report: Reducing Internet Latency 2013,” *ACM SIGCOMM CCR*, vol. 44, no. 2, Apr. 2014, pp. 80–86.

[2] A. Singla *et al.*, “The Internet at The Speed of Light,” *ACM Hotnets 2014*, Oct. 2014, Los Angeles, CA, pp. 1–7.

[3] G. Fettweis *et al.*, “The Tactile Internet,” ITU-T Technology Watch Report, Aug. 2014, pp. 1–24.

[4] T.H. Szymanski, “An Ultra Low Latency Guaranteed-Rate Internet for Cloud Services,” *IEEE Trans. Networking*, vol. 24, no. 1, Feb. 2016, pp. 123–36.

[5] N. Finn and P. Thubert, “Deterministic Networking Problem Statement (04),” IETF Internet Draft, Standards Track, Oct. 19, 2015, pp. 1–17.

[6] S. Shah and P. Thubert, “Deterministic Forwarding PHB (04),” IETF Internet Draft, Aug. 30, 2015, pp. 1–8.

[7] S. Iyer, R. R. Kompella, and N. Mckeown, “Designing Packet Buffers for Router Linecards,” *IEEE Trans. Networking*, vol. 16, no. 3, June 2008, pp. 705–17.

[8] V. Anantharam *et al.*, “Achieving 100% Throughput in an Input Queued Switch,” *IEEE Trans. Commun.*, vol. 47, no. 8, 1999, pp. 1260–67.

[9] W.J. Chen, C-S. Chang, and H-Y. Huang, “Birkhoff-von Neumann Input Buffered Crossbar Switches for Guaranteed-Rate Services,” *IEEE Trans. Commun.*, vol. 49, no. 7, July 2001, pp. 1145–47.

[10] I. Keslassy *et al.*, “On Guaranteed Smooth Scheduling for Input-Queued Switches,” *IEEE/ACM Trans. Networking*, vol. 13, no. 6, Dec. 2005, pp. 1364–75.

[11] IEEE 802 Tutorial, “Deterministic Ethernet: 802.1 Standards for Real-Time Process Control, Industrial Automation, and Vehicular Networks,” Nov. 12, 2012, pp. 1–72.

[12] D. Dujovne *et al.*, “6TiSCH: Deterministic IP-Enabled Industrial Internet (of Things),” *IEEE Commun. Mag.*, vol. 52, no. 12, Dec. 2014, pp. 36–41.

[13] P. Wetterwald and J. Raymond, “Deterministic Networking Utilities Requirements,” IETF Internet Draft, June 30, 2015, pp. 1–26.

[14] T. H. Szymanski, “Crossbar Switch and Recursive Scheduling,” U.S. Patent 9042380B2, issued May 26, 2015, pp. 1–36.

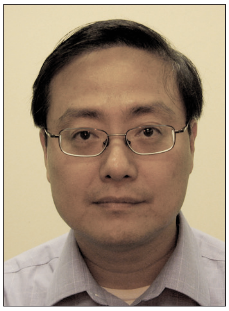
[15] T. H. Szymanski, “Max-Flow Min-Cost Routing in a Future Internet with Improved QoS Guarantees,” *IEEE Trans. Commun.*, vol. 61, no. 4, Apr. 2013, pp. 1485–97.

Biography

Ted H. Szymanski (teds@mcmaster.ca) completed his Ph.D. degree at the University of Toronto. From 2001 to 2011, he held the Bell Canada Chair in Data Communications at McMaster University. Previously, he was a professor at Columbia University and its Center for Telecommunications Research, and McGill University and the Canadian Institute for Telecommunications Research. He participated in a 10-year research program within the Networks of Centers of Excellence of Canada, which demonstrated a free-space intelligent optical backplane using photonic packet-switches with about 1000 optical channels. Contributors included Nortel Networks (Ericsson), Newbridge Networks (Alcatel), Lockheed-Martin/Sanders, and McGill, McMaster, Toronto, and Heriot-Watt Universities. His group also demonstrated the first FPGA with optical IO, using the U.S. DARPA/Lucent/Coop smart-pixel foundry service. His interests include security, energy efficiency, deterministic communications, and the industrial Internet of Things.

A speed-of-light deterministic core network can have a profound impact on virtually all consumer services such as multimedia distribution, e-Commerce, and cloud computing or gaming. It can also pay for itself quickly, due to its significantly improved utilization and energy efficiency.

Automotive Networking and Applications



Wai Chen

Luca Delgrossi

Timo Kosch

Tadao Saito

In this 17th issue of the Automotive Networking and Applications Series, we are pleased to present two articles that address:

- Enhancing a current in-car networking technology to meet the rapidly rising needs of the onboard electronic components
- A biologically inspired self-organizing traffic control scheme, based on vehicle-to-vehicle (V2V) communications, that expedites emergency vehicles (EVs) through urban traffic intersections

An important urban safety issue is how to support emergency vehicles (EVs), such as ambulances or fire trucks, to expedite the EVs through the urban traffic to reach their destinations as fast as possible. This support of EVs becomes an especially interesting challenge when vehicles use vehicle-to-vehicle communications to self-organize themselves to pass through traffic intersections. The first article, "A Self-Organizing Network Approach to Priority Management at Intersections" by O.K. Tonguz and W. Viriyasitavat, proposes a biologically inspired self-organizing traffic control scheme that expedites the EVs through traffic intersections in urban areas. In the article, the authors first outline their motivation of creating a "green-wave" effect for the EVs and then give an overview of the self-organized traffic control paradigm. The article then describes in detail the proposed algorithm for enabling prioritized intersection control, where vehicles communicate among themselves using V2V communications to resolve potential conflicts at the intersections and assign (high) priority to EVs. The authors then present extensive simulation results to demonstrate the efficacy of their proposed scheme in reducing the travel time of EVs through urban traffic while exacting only a negligible adverse effect on non-emergency vehicles. The article also discusses some open challenges related to the priority management of EVs.

As the number and type of electronic components in cars increase, the need for ever higher data transmissions among these components within the cars also rises rapidly. The controller area network (CAN) is the most widely deployed in-car networking technology, which is characterized by its simplicity and robust performance, albeit at somewhat low data rates (e.g., up to 1 Mb/s for CAN 2.0, or up to 16 Mb/s for CAN-FD) that cannot match other networking technologies such as Ethernet or optical fibers. Therefore, how to expand the applications of the CAN standard beyond its wide adoption in control components to various in-car electronic components that require higher data rates has become an interesting and timely challenge. The second article, "High-Speed CAN Transmission Scheme Supporting Data Rate over 100 Mb/s" by S. Kang *et al.*, proposes a new scheme to increase the data rate of the CAN network to over 100 Mb/s, while remaining seamlessly compatible with the existing CAN network protocol. In their article, the authors first review major in-car network standards including LIN, CAN/CAN-FD, FlexRay, Automotive Ethernet, and MOST. The authors then discuss the CAN network in greater detail, and propose and describe in detail their new scheme for the high-speed CAN network. Through their simulation analysis, the authors show that their proposed scheme can provide higher data rate while keeping backward compatibility with the existing CAN standard. The authors conclude with a discussion of some open challenges related to further enchantments to CAN.

We thank all contributors who submitted manuscripts for this Series, as well as all the reviewers who helped with thoughtful and timely reviews. We thank Dr. Osman Gebizlioglu, Editor-in-Chief, for his support, guidance, and suggestions throughout the process of putting together this issue. We also thank the IEEE publication staff, particularly Ms. Peggy Kang and Ms. Jennifer Porcello, for their assistance and diligence in preparing the issue for publication.

Biographies

Wai Chen (waichen@ieee.org) received his B.S. degree from Zhejiang University, and M.S., M.Phil., and Ph.D. degrees from Columbia University, New York. He is chief scientist of the China Mobile Research Institute and general manager of the China Mobile Internet-of-Things Research Institute. Previously he was VPGD of ASTRI, Hong Kong, and chief scientist and director at Telcordia (formerly known as Bellcore), New Jersey.

Luca Delgrossi is manager of the Vehicle-Centric Communications Group at Mercedes-Benz Research & Development North America Inc., Palo Alto, California. He received his Ph.D. in computer science from the Technical University of Berlin, Germany. He served for many years as professor and associate director of the Centre for Research on the Applications of Telematics to Organizations and Society (CRATOS) of the Catholic University at Milan, Italy.

Timo Kosch is a team manager for BMW Group Research and Technology where he is responsible for projects on distributed information systems, including cooperative systems for active safety and automotive IT security. He studied computer science and economics at Darmstadt University of Technology and the University of British Columbia in Vancouver. He received his Ph.D. from the computer science faculty of the Munich University of Technology.

Tadao Saito [LF] received his Ph.D. degree in electronics from the University of Tokyo. He is a professor emeritus at the University of Tokyo. He was chief scientist and CTO of Toyota InfoTechnology Center. He is chairman of the Ubiquitous Networking Forum of Japan and chairman of the Next Generation IP Network Promotion Forum of Japan. He has published eight books on electronics, computers, and digital communications. He is a Fellow of IEICE of Japan.

A Self-Organizing Network Approach to Priority Management at Intersections

Ozan K. Tonguz and Wantanee Viriyasitavat

ABSTRACT

Prior work has shown that a biologically inspired approach can solve some of the fundamental transportation problems in urban areas. As one instance of this approach, it was shown that vehicles equipped with dedicated short-range communications (DSRC) radios can manage traffic in urban areas in a completely self-organized manner similar to self-organizing biological systems (e.g., ants, birds, and fish). This scheme is known as virtual traffic lights, and its success is enabled by the design of local rules which allow vehicles approaching an intersection to resolve the ensuing conflict in a seamless and self-organized manner without the need for any infrastructure. One important safety issue in urban traffic is how to manage the presence of emergency vehicles such as ambulances and fire trucks. In this article, it is shown that by designing a different set of local rules, one can give priority to emergency vehicles at every intersection, thus expediting their response times.

INTRODUCTION

Among the different safety applications of vehicular ad hoc networks (VANETs), all except the post crash notification (PCN) application are designed to *prevent* accidents; the PCN application is a passive safety application which aims at disseminating safety information to intended drivers so that they are informed about an accident in a timely manner and make rerouting decisions when necessary. Therefore, the main objective of this application is to direct normal traffic around an accident area, but *not* to assist an emergency response team to reach the scene of an accident.

Using the self-organized traffic control paradigm first proposed in [1], we propose the use of dedicated short-range communications (DSRC) and VANET technologies to enable an *active and post-incident* safety application that aims at facilitating and prioritizing the motion of emergency vehicles (EVs) through traffic in urban areas. While law enforcement to facilitate such EV motion is already in place (i.e., vehicles move over to give way to approaching EVs), further improvement in emergency response time is extremely critical, especially in fire and health-related incidents. To put things into perspective, Fig. 1 depicts the generalized flashover curve for residential constructions. A flashover occurs at the stage of an ensuing fire at which all surfaces and objects within a space have been heated to their ignition temperature, and flame breaks out almost at once over the surface of all objects in the space; hence, it is the most dangerous part of a fire for firefighters. As shown in the figure, a reduction in response time (i.e., the time that lapses after the fire is detected and reported until the time the firefighters arrive at the scene) is crucial. Similarly, in the case of health-related incidents, it has been shown that the chance of survival drops by roughly 10 percent for every minute a patient stays in cardiac arrest (see, e.g., http://www.americanheart.org).

Reducing the emergency response time by minutes or even seconds is therefore crucial in an emergency situation. The design of the proposed system is based on a concept similar to what is currently being used today; that is, assigning highest priority to EVs. However, instead of allowing the EVs to pass through intersections without obeying traffic lights, our system aims to provide a "green-wave" effect for the EVs. Green wave phenomena occur when a series of traffic lights are synchronized so that the number of times a car needs to stop at intersections is minimized. This allows continuous traffic flow and significantly improves traffic flows for EVs at intersections. By having always-green signals displayed to the EVs, the proposed system allows the EVs to move at a faster speed and also avoids EV-involved accidents (e.g., each year in the United States, 80 EV crashes occur that involve fatalities [2], and, on average, the EV travel time can be reduced by at least 26 percent [3]). In addition, it is reported in [4] that a major portion of EV crashes take place at intersections, and among these, more than 25 percent have been found to occur at signalized intersections — vehicles approaching a green signal cannot see an EV approaching from the intersecting roadway because of line-of-sight problems due to nearby buildings, vegetation, or hills [2]. By presenting green and red signals to the EV and other non-EV vehicles, respectively, these EV crashes at intersections could be avoided.

In addition, it is worth pointing out that since the proposed scheme enables traffic control (if necessary) at every intersection, the EV-involved accidents that have been found to occur

One important safety issue in urban traffic is how to manage the presence of emergency vehicles such as ambulances and fire trucks. The authors show that by designing a different set of local rules, one can give priority to emergency vehicles at every intersection, thus expediting their response times.

Ozan K. Tonguz is with Carnegie Mellon University; Wantanee Viriyasitavat is with Mahidol University and Norwegian University of Science and Technology.

0163-6804/16/$25.00 © 2016 IEEE

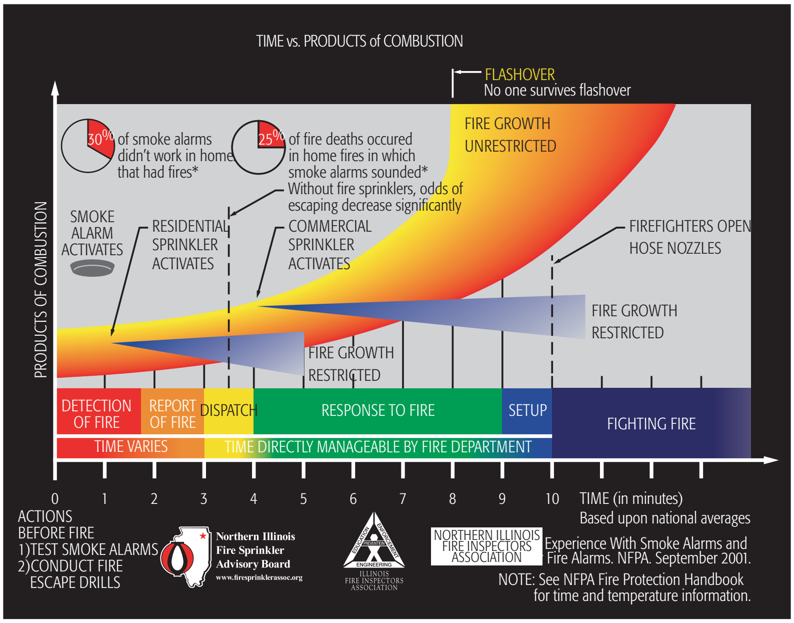


Figure 1. The flashover curve and the evolution of fire with time in a building. Chance of survival is substantially decreased after the flashover point (picture taken from http://firesprinklerassoc.com/wp-content/uploads/2014/02/newflashoverchart-1024x791.jpg).

at non-signalized intersections could also be prevented. As a result, benefits of the proposed scheme are *ubiquitous*: at signalized and non-signalized intersections.

Self-Organized Traffic Control

The premise of the priority management scheme proposed in this article is the self-organized traffic control paradigm first proposed in [1]. It was shown later that this self-organized traffic control paradigm is an instance or example of biologically inspired solutions to several fundamental transportation problems [5]. In the proposed self-organized traffic control paradigm, vehicles communicate among themselves (i.e., in an ad hoc manner without any help from infrastructure) to resolve conflicts at intersections and determine who should cross the intersections first (i.e., they establish the right of way). Unlike other self-organized traffic control schemes, which rely on some centralized infrastructure [6], the proposed scheme operates in a distributed manner under the assumption that each vehicle periodically broadcasts hello messages to announce its presence, current position, and velocity to other nearby vehicles. A vehicle can therefore construct a local map and determine if there is an ensuing conflict at the intersection it is about to approach. In situations where a conflict is detected, vehicles involved in the conflict perform the following three steps:

- Leader election process
- Generation of traffic light information
- Handover

Based on the beaconing mechanism of DSRC technology, cluster leaders approaching an intersection can communicate and choose one cluster leader to serve as a "virtual traffic light" (VTL). The leader election process is based on safety considerations as well as other considerations. In general, the cluster leader (among the four cluster leaders) farthest from the intersection is chosen as the VTL. Upon acknowledgment of the VTL by the other cluster leaders, the VTL broadcasts a red light to its approach and a green light to the orthogonal direction. After a fixed duration (e.g., 45 s) and/or possibly based on other criteria (e.g., the number of vehicles in each approach), the VTL responsibility can be handed over to a cluster leader in the orthogonal direction. More details about the principle of operation of the VTL scheme can be found in [1].

It has been shown by extensive simulations that the aforementioned traffic control scheme (i.e., the VTL system) could provide up to 60 percent improvement in traffic flow [1]. Such a significant improvement is due to two reasons: i) VTL can render traffic control truly ubiquitous compared to only 20 percent of intersections that are currently equipped with traffic lights; and ii) VTL reduces the "dead period" of intersections (i.e., unnecessary red lights when a green light is given to a road with no vehicles).

It should be noted that the above VTL system operates based on the following assumptions [1]:

- All vehicles are equipped with DSRC radios.
- All vehicles share the same digital map and positioning system device that has lane-level accuracy.
- The RF propagation problems such as obstructions due to buildings at the corners of intersections do not disrupt the necessary vehicle-to-vehicle communication for electing a leader that will serve as a VTL [7].
- Other communications problems due to collision of transmitted packets or beacon messages by vehicles are not severe.

Proposed Algorithm for Enabling Priority Intersection Control

It has been shown in several previous studies that VTL is a biologically inspired self-organizing network approach to urban traffic control, and this scheme ultimately depends on designing local rules at intersections for deciding the right of way between competing flows. By obeying these local rules, the flow rate can be increased by 30–60 percent, which is a significant improvement in mitigating traffic congestion and reducing the commute time during rush hours. In this article, we show that by designing new local rules for intersections, it is possible to perform yet another important functionality: giving priority to emergency vehicles including ambulances, fire trucks, and so on. This expedites the movement of emergency vehicles, which, in turn, can save many lives.

It is well known that the local rules used by self-organizing biological systems in nature for different functionalities are different. While a certain set of local rules might be used for foraging, a different set of local rules might be used by social insect colonies for protecting themselves against predators, and yet another set of rules might be used to cope with drastic changes in the environment (e.g., in terms of temperature). Depending on the nature of the function to be performed, these local rules might necessitate using priority rules. As an example, it is interesting to note that priority rules (which are functions of factors such as size, direction of movement, and whether or not they carry a load) have been used in a number of ant species for deciding the right of way [8, 9]. For instance, in the leaf-cutting ant *Atta columbica*, higher priority is given to the inbound laden ants as the outbound ants give way to inbound laden ants in 80 percent of their encounters, possibly due to the fact that the inbound ants are carrying leaves or food for the colony. Inspired by such observations, we propose a self-organizing network solution to one of the major problems in emergency response management [5]; that is, facilitating and expediting the motion of EVs (or high-priority vehicles) through traffic and/or congestion in urban areas. By detecting the presence of an EV, the proposed scheme, VTL with priority intersection control (VTL-PIC), assigns priority (i.e., gives right of way) to the road or approach on which the EV travels. To enable the priority scheme at intersections, *two new mechanisms* (i.e., local rules) are designed [10].

Detection of an EV When It Approaches and Leaves an Intersection: It is clear that detection of an EV is a critical component of the proposed VTL-PIC scheme. In our proposed solution, upon approaching an intersection, the EV periodically broadcasts a PIC request message to announce its presence and demand priority until it receives a PIC grant message from a vehicle that is leading the intersection (i.e., the intersection leader). Note that in addition to the PIC request message, the intersection leader can detect the presence of the EV when it receives a hello message generated by the EV.

Besides PIC request messages, the EV is also required to inform the intersection leader upon leaving the intersection so that the intersection can now resume its normal operation for normal traffic management. A PIC clear message is used to handle such detection. When the EV crosses the conflict point (intersection), it periodically broadcasts a PIC clear message for a certain period of time. In the case when PIC clear messages are lost, the intersection leader can also detect the departure of the EV when it does not receive hello messages from the EV for a certain period of time.

Priority Assignment Scheme: Once the presence of an EV is detected, phase layout configuration of the traffic signals of the intersection needs to be recomputed and broadcast to vehicles involved in the conflict at the intersection. While there are a number of algorithms that could be used for priority assignment, a simple scheme (i.e., the cluster in which the EV is traveling always gets the green signal) is used in our protocol to illustrate how priority intersection control could be used in conjunction with the VTL system.

Since the proposed scheme is an overlay scheme on top of VTL, the VTL-PIC will also share the same benefits as the VTL scheme. The benefits of the VTL-PIC protocol on the travel time of emergency vehicles are as follows.

Less Severe Traffic Congestion for EVs: Given the same amount of traffic, an EV in the VTL-PIC scheme encounters less severe traffic congestion compared to that found in typical scenarios with physical traffic lights. Our previous work has shown that because of more efficient use of intersections as a *resource* and the fact that the VTL renders traffic control ubiquitous (i.e., traffic control at every intersection), during rush hours traffic congestion takes place at a much later stage (i.e., more vehicles can enter the network before traffic congestion happens) compared to typical scenarios with physical traffic lights and identical traffic generation rate. As a result, vehicles and especially an EV reach their destination locations within a much shorter time duration when the VTL scheme is employed. Furthermore, when traffic congestion is inevitable (i.e., generated traffic exceeds the capacity of the road network), the VTL scheme can resolve the congestion situation much more quickly; hence, the travel time of the EV is substantially reduced.

Lower Travel Time for EVs: Creation of a green wave effect allows EVs to travel at higher speeds such that they do not need to slow down when approaching intersections. Since the VTL-PIC scheme always assigns green signals to the road on which the EVs are traveling, the EVs will encounter green-wave phenomena as they pass through intersections (i.e., consecutive traffic lights are coordinated and present progressive green displays to the EVs). This results in a higher traveling speed of EVs compared to conventional operation where EVs have to slow down significantly as they see red signals when approaching intersections. By assigning higher priority to the roads on which EVs travel, VTL-PIC could clear up an EV's route by also giving higher priority to vehicles that travel in front of the EV to pass the intersections.

Potential Prevention of Crashes: The proposed scheme can prevent potential EV crashes or accidents that take place at intersections. By presenting a green signal in the EV's direction (approach), other vehicles that approach from different directions are presented with red signals, and hence are prepared to stop and give right of way to the EV. Note that because of ubiquitous traffic control rendered by the VTL-PIC scheme, it is expected that the proposed protocol could prevent a significant number of EV crashes (i.e., more than 25 percent [2]), thus making *both signalized* and *non-signalized* intersections safer for both emergency and regular (non-emergency) vehicles.

Principle of Operation

According to the scheme described above, a vehicle at an intersection must belong to one of three different categories: an EV, a non-EV vehicle that is leading the intersection, and a non-EV vehicle that is not leading the intersection. Figures 2a and 2b depict the flow diagrams of the algorithms used by an EV and a non-EV vehicle upon approaching an intersection, respectively.

Upon approaching an intersection, an EV determines if there is already a VTL set up for the intersection by passively listening to the VTL message broadcast by the leader. In the case when no VTL exists and no conflict is detected at the intersection, the EV can pass through the intersection with no additional communication. In the case where a conflict is detected or

When the EV crosses the conflict point (intersection), it periodically broadcasts a PIC clear message for a certain period of time. In the case when PIC clear messages are lost, the intersection leader can also detect the departure of the EV when it does not receive hello messages from the EV for a certain period of time.

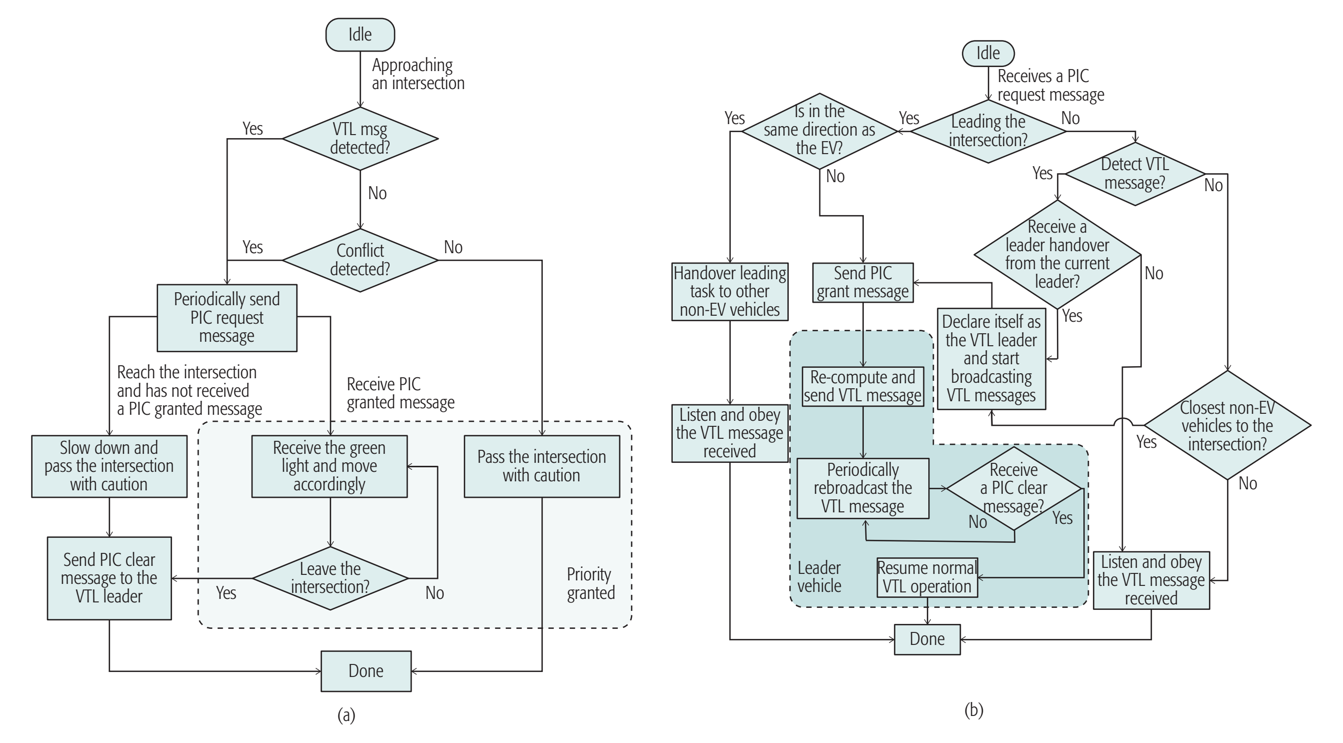

Idle
Approaching an intersection
VTL msg detected?
Yes
No
Conflict detected?
Yes
No
Periodically send PIC request message
Reach the intersection and has not received a PIC granted message
Receive PIC granted message
Slow down and pass the intersection with caution
Receive the green light and move accordingly
Pass the intersection with caution
Send PIC clear message to the VTL leader
Leave the intersection?
Yes
No
Priority granted
Done
(a)
Idle
Receives a PIC request message
Leading the intersection?
Yes
No
Is in the same direction as the EV?
Yes
No
Detect VTL message?
Yes
No
Receive a leader handover from the current leader?
No
Yes
Handover leading task to other non-EV vehicles
Send PIC grant message
Declare itself as the VTL leader and start broadcasting VTL messages
Re-compute and send VTL message
Listen and obey the VTL message received
Closest non-EV vehicles to the intersection?
Yes
No
Periodically rebroadcast the VTL message
Receive a PIC clear message?
No
Yes
Leader vehicle
Resume normal VTL operation
Listen and obey the VTL message received
Done
(b)

Figure 2. Flow diagram describing the principle of operation of an emergency vehicle and a non-emergency vehicle when they approach an intersection: a) emergency vehicle; b) non-emergency vehicle.

a VTL is already set up, the EV announces its presence and requests priority for right of way at the intersection by sending a *PIC request* message to the intersection leader. Note that in the case where the leader has not been elected, the closest vehicle to the intersection that travels in the orthogonal direction to that of the EV is automatically chosen as the leader (Fig. 2b). The PIC request is periodically transmitted until the EV receives a *PIC grant* message sent from the leader to acknowledge the presence and granted priority to the EV. In the unlikely case where the EV reaches the intersection and has not received a PIC grant message, the EV resorts back to the conventional procedure; that is, it slows down and watches for other vehicles before it crosses the intersection. As soon as the EV leaves the intersection, it broadcasts a *PIC clear* message to the leader to *release* the intersection for normal traffic use.

A flow diagram of the algorithm used for non-EV vehicles is shown in Fig. 2b. When the leader vehicle receives a PIC request message (or a hello message) sent from the approaching EV, the leader determines if it should continue to lead the intersection. In other words, in the case when the leader is traveling in the same approach in front of the EV and blocking the EV's movement, the leader hands its leading task over to other vehicles. Otherwise, the leader that does not block the movement of the EV continues to lead the intersection, and replies to a PIC request message with a PIC grant message. To permit the EV to pass through the intersection, the leader recomputes the phase layout of the traffic signals and communicates the new configuration to all vehicles in the intersection. Once the leader detects the EV leaving the intersection (either through the reception of a PIC clear message or several omissions of hello messages from the EV), the leader recomputes the traffic signal configuration to allow normal traffic management.

Upon receiving a PIC request message from an EV, other non-EV vehicles that do not assume the leading task could become the leader in one of the following two circumstances:

- If there is no VTL currently set up for the intersection, the vehicle elects itself as the intersection leader if it is the closest non-EV vehicle to the intersection (in the orthogonal direction).
- If a VTL has been set up for the intersection and the vehicle receives a *handover* message from the current leader, it assumes the leading task and becomes the new leader.

A vehicle that assumes the leading task (or leading role) because of one of the above scenarios needs to transmit a PIC grant message to the EV, compute the corresponding phase layout of traffic signals, and broadcast the traffic light message to all vehicles. Similar to the VTL scheme, other vehicles remain as passive nodes (i.e., they listen and obey the traffic light message they receive).

Priority Assignment Scheme

As mentioned previously, there are a number of different algorithms and schemes one could use to implement a priority intersection control scheme. Since the focus of this article is not on finding an optimal priority assignment scheme, a simple scheme is used in the proposed VTL-PIC protocol as an illustrative example to show how one could incorporate priority control into a self-organized traffic control system such as VTL. To elaborate on this, in the VTL-PIC scheme

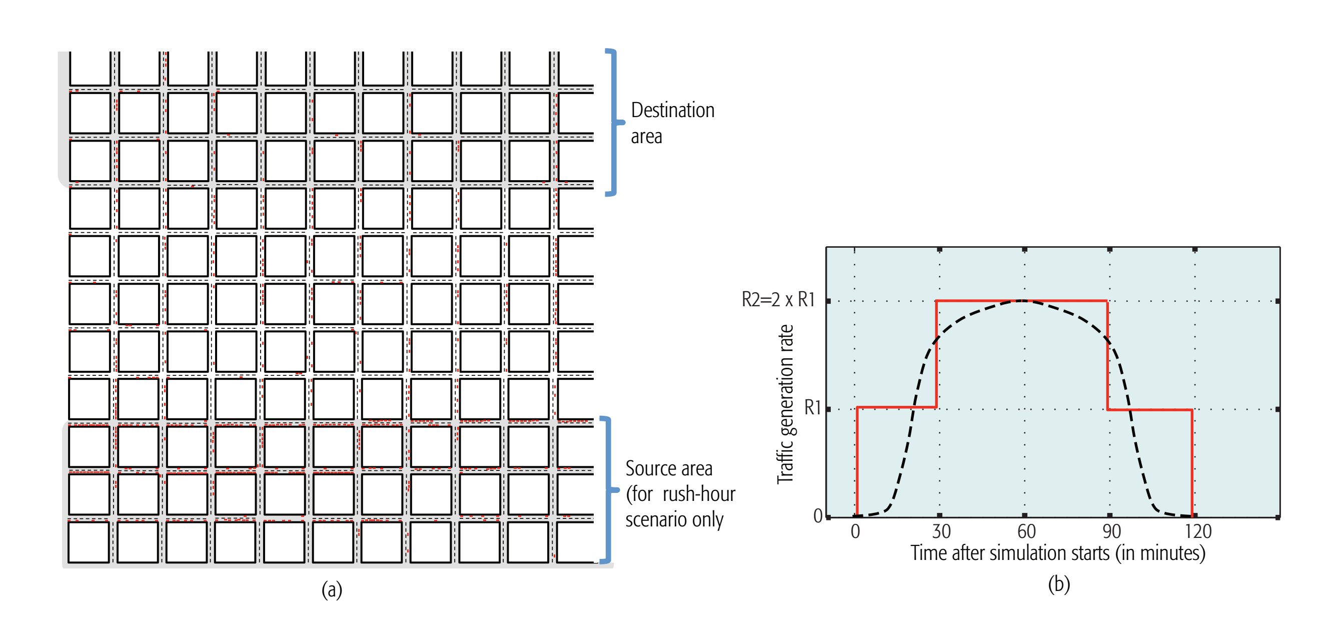

Destination area
Source area (for rush-hour scenario only
(a)
R2=2 x R1
R1
0
Traffic generation rate
0
30
60
90
120
Time after simulation starts (in minutes)
(b)

Figure 3. Left: A 10 × 10 Manhattan grid topology with 125-m block length is used in the simulations. The bottom 3 × 10 area is the source area where vehicles are injected into the network in the simulations. Small red dots represent vehicles. Right: Traffic generation pattern used in the simulations. While the black dotted line shows the realistic traffic generation rate, the staircase in red color shows the approximation used in the simulations: a) network topology; b) traffic generation patterns.

described in this article, an intersection leader always gives right of way (i.e., green light) to the road on which an EV is traveling. "Always-green" configuration continues until the EV has left the intersection, and normal operation of the VTL-PIC is then resumed. It is an interesting subject for future study to determine the optimal priority scheme for priority vehicles.

Simulation Setting and Results

Simulation Setting

In order to evaluate the proposed VTL-PIC protocol, we resort to the SUMO traffic mobility simulator, an open source microscopic simulator developed by the Institute of Transportation Systems at the German Aerospace Center [11]. A 10 × 10 Manhattan grid network topology is assumed in the simulations with 125-m block length.

The traffic generation pattern used in the simulations is depicted in Fig. 3b where the traffic generation rate (e.g., R_1 and R_2 [veh/h]) varies based on different time windows during rush hour, number of total vehicles injected into the simulations, N. The step function shown in Fig. 3b is used to capture the traffic behavior during rush hour; there is a first wave of commuters who try to enter/leave the city sooner to avoid traffic jams, followed by the period when most commuters enter/leave; and finally, another wave of the remaining vehicles. Hence, in this article the relationship between these three parameters is assumed to be

$$R_1 = \frac{N}{3}, \quad R_2 = 2R_1 = \frac{2N}{3}$$

One EV is artificially added into the simulation at t = 5400 s (i.e., 90 min after the simulation starts). The EV starts from the center of the source area to its destination in the top right of the network. All vehicles including the EV are assumed to be equipped with a GPS system and DSRC radios with a transmission range of 200 m. We assume that there is no packet loss in the network; that is, all packets sent are correctly received at the receiver(s). Travel time of the EV and non-EV vehicles are collected from the simulations conducted and reported in the next subsection.

Three different traffic control schemes are implemented and evaluated:

- A baseline scheme where only physical traffic lights (TLs) are used at intersections, and an EV does not receive any priority at intersections. In this scheme, an EV is treated as a non-EV vehicle. This assumption is valid in a heavily congested urban scenario; vehicles cannot move to the side to give way to the EV.
- A VTL scheme where the VTL paradigm is used as the traffic control mechanism at intersections; however, it does not give priority to the EV.
- A VTL-PIC scheme where both the VTL and the priority scheme are implemented.

Simulation Results

Rush Hour Scenario: To simulate the traffic pattern during rush hours, vehicles in the simulator are assumed to start from their origination location (source area) in the 3 × 10 source area located at the bottom of Fig. 3a to their destination area in the topmost portion of the network.

Figures 4a and 4b show the simulation results in terms of travel time of the EV and non-EV vehicles as a function of total number of vehicles generated, respectively. Observe that the travel time of both types of vehicles decreases when the VTL system is in place, and the proposed VTL-

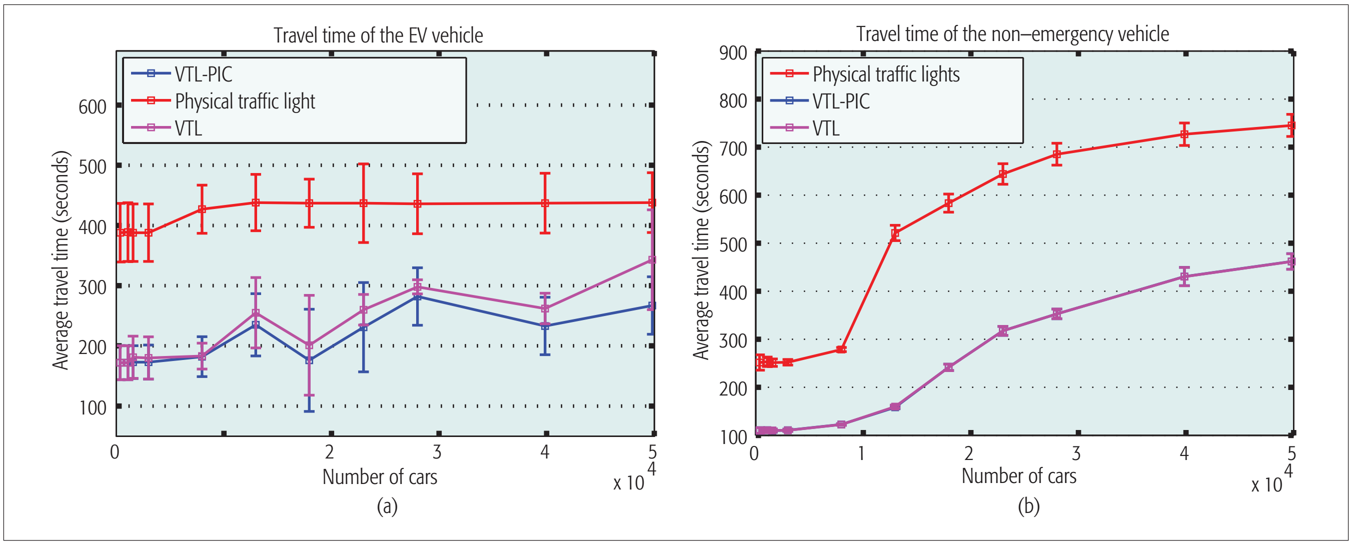


Figure 4. Average travel time of emergency and non-emergency vehicles in a scenario where an emergency vehicle is inserted into the network at $t = 90$ min after the onset of rush hour. The results are plotted with 95 percent confidence interval. Note that the confidence interval for the non-emergency vehicles is very small. Observe that in b) the results for VTL and VTL-PIC are almost the same: a) emergency vehicle; b) non-emergency vehicle.

PIC protocol further decreases the travel time of the EV vehicle. Figure 4a, for example, shows that with the new VTL-PIC scheme, the travel time of an EV can be reduced from 430 s (in the case with the TL scheme) to a range of 180–250 s, which is quite significant as such a 3- or 4-min reduction in travel time could save many lives in emergency response applications. It is worth pointing out that despite the enforced EV priority, VTL-PIC has little or no adverse effect on the travel time of non-EV vehicles (Fig. 4b).

Figure 5 presents in detail the travel time of the EV for each intersection it crosses. Note that based on the pre-specified route depicted in Fig. 3a, the EV passes three intersections before it leaves the source area and 12 more intersections outside the source area before it reaches its destination. As a result, the EV always encounters conflicts as it arrives at the first three intersections, but not afterward. This is because traffic density outside the source area is very low; thus, it is unlikely to encounter conflicts at the intersections outside the source area. The reduction in the travel time of the EV is therefore gained from the first three intersections. Furthermore, the advantage of the VTL-PIC scheme becomes more pronounced as the number of vehicles increases (Fig. 5, bottom). This confirms our intuition: VTL-PIC outperforms VTL only at the always conflicting intersections and when there are larger numbers of vehicles in the simulations, thus leading to a higher level of conflicts at intersections. It is important to note that the VTL-PIC protocol creates the green wave phenomena for the EV; that is, the time it takes for the EV to cross the intersection(s) is minimal, as shown in Fig. 5. As shown later, when the traffic pattern of non-EVs is uniformly distributed (instead of the dominant northbound traffic), the benefits of VTL-PIC will be more pronounced.

Lunchtime Scenario: In contrast to the 3 × 10 source area used in obtaining the previous results to simulate the traffic pattern during rush hour, the entire 10 × 10 network shown in Fig. 3a is used as the source area for lunchtime scenarios. All non-EVs have random start and end locations. Similar to the previous case, an EV starts from the center of the 3 × 10 area in the bottom part of the network to a destination in the top-right portion of the network (Fig. 3a).

Figures 6a and 6b depict the average travel time of the EVs and non-EVs, respectively. Observe that, compared to the physical TL scheme currently used, the average travel time of EVs can be reduced by up to 5 min with the VTL/VTL-PIC scheme, which is quite significant. Furthermore, compared to the VTL scheme, up to 45 s of travel time for the EV can be saved with the proposed VTL-PIC scheme. It is worth pointing out that given the same number of vehicles in simulations, benefits of VTL-PIC over VTL become more obvious (i.e, in a scenario with 7000 vehicles, the VTL-PIC scheme could save an additional 45 s of an EV's travel time vs. less than 5 s additional saving time obtained in the rush hour scenario). This larger benefit is due to the fact that the EV experiences conflicts at all of the 16 intersections it crosses in this scenario (as opposed to only 3 intersections observed in the rush hour scenarios). Note that the benefit in terms of travel time of an EV largely depends on the number of congested intersections; hence, the expected benefit will increase considerably when a larger urban area is considered.

Discussion

Since the proposed priority management scheme uses the existing VTL system described in [1] as its premise, a number of issues related to wide adoption of the VTL system are also relevant for the VTL-PIC scheme. More specifically, the proposed priority management scheme in this article assumes the full adoption of the VTL scheme by all the vehicles in an urban area. While such a fundamental paradigm shift can be brought about by the Departments of Transportation (DoTs) of different countries based on new legislation, our ongoing work shows that it might be possi-

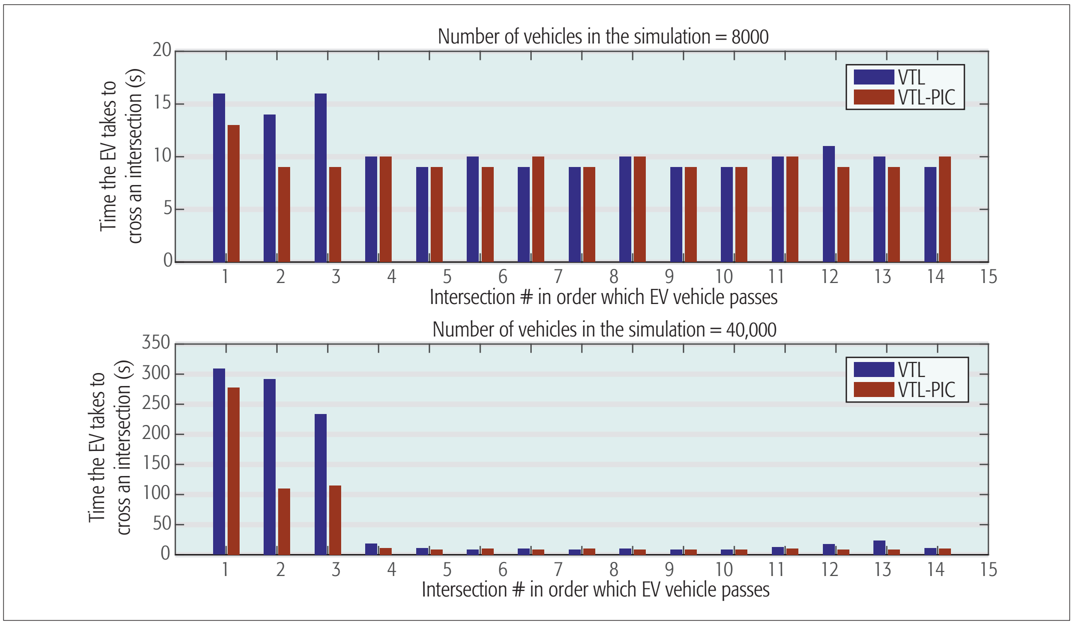


Figure 5. Average travel time of an EV as it passes through 15 intersections before it reaches its destination, 3 of which are intersections in the source area and always have conflicts. Note that the results presented here are extracted from one of many simulation runs; the negligible discrepancy observed in the top figure (i.e., the EV crossing time at intersections 7 and 15 are slightly higher in the VTL-PIC scheme) is the result of simulation artifacts.

Only VTL vehicles are allowed to travel on these streets while non-VTL vehicles can still utilize other streets. This concept thus allows the co-existence of VTL and non-VTL vehicles in the same road network and also implies that the proposed priority management scheme might be applicable on those routes where VTL is enforced by the local DoT's.

ble to deploy VTL partially on designated routes only; for example, during rush hours using an approach similar to the high occupancy vehicle (HOV) concept [12]. In this HOV-like concept, several streets are exclusively reserved for VTL vehicles and intersections on these streets adopt VTL as the intersection control management [12]. Only VTL vehicles are allowed to travel on these streets while non-VTL vehicles can still utilize other streets. This concept thus allows the coexistence of VTL and non-VTL vehicles in the same road network, and also implies that the proposed priority management scheme might be applicable on those routes where VTL is enforced by local DoTs. The same scheme can also be applied to the VTL-PIC scheme proposed in this article in scenarios where the technology penetration rate is less than 100 percent. Similar to the results shown in [12], depending on the route that needs to be taken by an EV, the benefit of the proposed VTL-PIC scheme might significantly degrade when the penetration rate is less than 100 percent.

In addition, the successful operation of the proposed scheme also assumes the right driver interface, which involves some kind of display unit on the dashboard of every vehicle that will interface with the DSRC unit. The fault-tolerant and security aspects of the proposed scheme are also major considerations for ensuring that, even in the worst case scenarios (e.g., in the event of high packet loss and inaccurate GPS information), the proposed scheme does not lead to accidents. Such provably fail-safe operation must be guaranteed [13]. Finally, the proposed scheme has to have the capability of dealing with pedestrians and cyclists who might not be equipped with DSRC radios. An external representation for the VTL on the outside of the vehicle has been proposed to display traffic light information to non-equipped cars, pedestrians, and cyclists [14]. Note that this might be less of a problem, though, given that most pedestrians or cyclists will give way to the EVs once they hear an emergency siren.

It is also worth mentioning that the priority assignment scheme introduced in this article is only a proof-of-concept type of example that illustrates how the powerful biologically inspired self-organized traffic control paradigm could be used to facilitate a safety-critical operation. As future work, it would be interesting to extend the proposed priority management scheme where several (instead of a single) consecutive intersections are coordinated to further increase the traffic efficiency. In addition, the same priority scheme could also be used to prioritize mass transit vehicles at intersections (buses, etc.) during rush hours, which, in turn, could reduce the commute time of urban workers during rush hours [15].

Related Work

Existing approaches to priority management at intersections for emergency, municipal, and mass transit buses are usually known as EV preemption (EVP) systems, and can be categorized into two main categories:

Centralized Control Systems

The most straightforward way of implementing an EVP mechanism is to employ a centralized control system. Examples of such systems include Global Traffic Technologies (GTT)'s Opticom™ Central Management Software (http://www.gtt.com) and GERTRUDE (http://www.gertrude.fr). In GTT's commercial product, arrival of EVs at an intersection is recognized by the traffic signal controller through light, sound, or radio waves. Once detected, the centralized urban traffic control software decides if signal preemption is war-

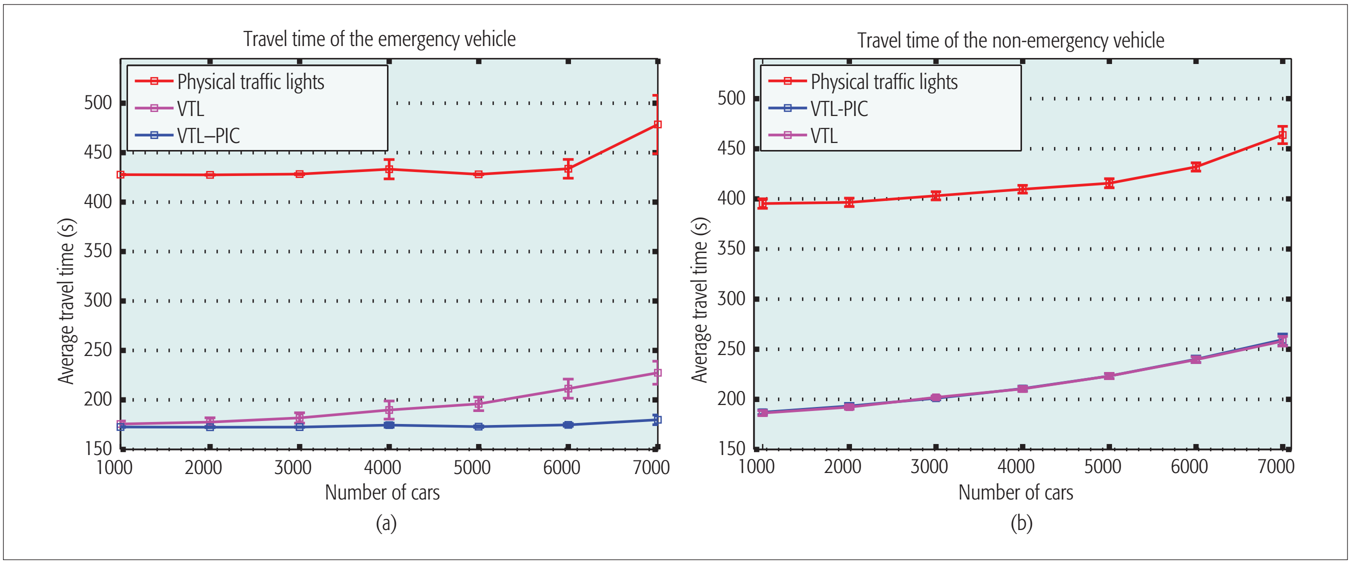


Figure 6. Average travel time (in seconds) of the emergency and non-emergency vehicles in the lunchtime scenario as a function of number of vehicles in the network where the emergency vehicle is inserted into the network at $t = 90$ min. The results are plotted with 95 percent confidence interval. Note that the confidence interval for the non-emergency vehicles is very small. Again, observe that the results of VTL and VTL-PIC are almost the same in Fig. 6b: a) emergency vehicle; b) non-emergency vehicle.

ranted and, if necessary, interrupts the normal green-yellow-and-red cycle to change the light to green for the EV. GTT's Optimal Product has been implemented in Olathe, Kansas, and Savannah, Georgia, and it has been shown that it creates efficiency in motion for the EVP, which allows first responders to arrive at an emergency scene faster.

Intersection-Based Systems

These systems can operate without a backbone network connecting all intersections to a central control center; hence, they offer a more scalable solution than the centralized system. Disadvantages of this approach are:

- Higher installation, operation, and maintenance cost per intersection
- Lower efficiency in EVP due to lack of coordination between consecutive intersections

Similar to the above approach, as an EV approaches a traffic signal, a light-, radio-, or sound-based detection mechanism triggers the traffic signal controller installed at the signal pole and/or arm to adjust the traffic light pattern. EMTRAC systems (http://www.emtracsystems.com) and E-ViEWs Safety Systems are the two most prevalent solutions. In EMTRAC, a 30–60 s reduction in travel time has been reported when the system is used in the mass transportation system in Santa Clara, California. However, the EVs and the intersections need special equipment that cost $3000 per vehicle and $10,000 per intersection.

E-ViEWs Safety Systems, partnered with NASA's Jet Propulsion Laboratory, has launched several technologies for emergency intelligent transportation systems (http://eviewsinc.com). Based on the wireless communications among EVs, dynamic message signs (DMSs), and intersection traffic controllers, road users are notified of approaching EV vehicles at signalized intersections; hence, the system provides faster emergency response times and greater safety on the roads.

While these systems have been shown to reduce response time of EVs, all of these schemes rely on some kind of infrastructure and require additional costly equipment to be installed; hence, coverage and benefits of such systems are fundamentally limited by the number of equipped intersections.

Conclusion

We have proposed a biologically inspired self-organized traffic control scheme that expedites emergency response operations (i.e., facilitates and expedites the movement of emergency vehicles through traffic in urban areas). In the proposed priority management scheme, vehicles communicate among themselves using vehicle-to-vehicle communications to resolve potential conflicts at intersections and determine a priority scheme that assigns the priority to a specific road or approach. The proposed priority scheme is based on local rules for vehicles approaching intersections, corresponding to the detection of the presence (and absence) of an emergency vehicle and rules that assign priority to emergency vehicles. Results of extensive simulations have shown that, with the proposed scheme, even in a relatively small topology (e.g., 10 × 10 Manhattan grid), an emergency vehicle is able to arrive at the scene of an accident up to 5 min earlier compared to a conventional physical traffic light system. Clearly, for emergency response operations this is quite significant. For bigger or denser cities, the reduction in travel time of emergency vehicles will be even larger. In addition, results also show that the proposed VTL-PIC protocol has a negligible adverse effect on the travel time of non-emergency vehicles.

Acknowledgment

This research has been supported in part by the T-SET University Transportation Center sponsored by the U.S. Department of Transportation under Grant No. DTRT12-G-UTC11.

References

[1] M. Ferreira *et al*., "Self-Organized Traffic Control," *Proc. ACM Int'l. Wksp. Vehicular Internetworking*, 2010, pp. 85–90.

[2] U.S. DOT (2003), "Fatality Analysis Reporting System (FARS) Web-Based Encyclopedia Queries for Emergency Use Crash Statistics," http://www-fars.nhtsa.dot.gov

[3] L. Faubion, "Emergency Vehicle Priority (EVP) Systems Reduce Response Time, Collision Avoidance, Fire Engineering," http://community.fireengineering.com/forum/topic/show?id=1219672 percent3ATopic percent3A317613, Apr. 2011.

[4] ITS, U.S. DoT, "Traffic Signal Preemption for Emergency Vehicles: A Cross-Cutting Study, Putting the 'First' in 'First Response,' " http://ntl.bts.gov/lib/jpodocs/repts_te/14097_files/14097.pdf, 2006.

[5] O. K. Tonguz, "Biologically Inspired Solutions to Fundamental Transportation Problems," Keynote, IEEE VNC, Jersey City, NJ, Dec. 2010.

[6] A. Leich, "Cooperative Self-Organizing System for Low Carbon Mobility at Low Penetration Rates," tech. rep., COLOMBO Project, Nov. 2015.

[7] T. Neudecker *et al.*, "Feasibility of Virtual Traffic Lights in Non-Line-Of-Sight Environments," *Proc. 7th ACM Int'l. Wksp. Vehicular Internetworking*, 2012, pp. 103–06.

[8] A. Dussutour *et al*., "Priority Rules Govern the Organization of Traffic on Foraging Trails under Crowding Conditions in the Leaf-Cutting Ant Atta Colombica," *J. Experimental Biology*, vol. 212, Feb. 2008, pp. 499–505.

[9] P. Casacci L *et al*., "Ant Pupae Employ Acoustics to Communicate Social Status in Their Colony's Hierarchy," *Current Biology*, vol. 23, Feb. 2013, pp. 323–27.

[10] W. Viriyasitavat and O. K. Tonguz, "Priority Management of Emergency Vehicles at Intersections Using Self-Organized Traffic Control," *Proc. IEEE VTC-Fall*, 2012, pp. 1–4.

[11] M. Behrisch *et al*., "Sumo - Simulation of Urban Mobility: An Overview," *Int'l. Conf. Advances in System Simulation*, Oct. 2011, pp. 63–68.

[12] O. K. Tonguz, W. Viriyasitavat, and J. Roldan, "Implementing Virtual Traffic Lights with Partial Penetration: A Game-Theoretical Approach," *IEEE Commun. Mag*., vol. 52, no. 12, Dec. 2014, pp. 173–85.

[13] J. Yapp and A. Kornecki, "Safety Analysis of Virtual Traffic Lights," *Proc. Int'l. Conf. Methods and Models in Automation and Robotics*, Aug. 2015, pp. 505–10.

[14] H. Conceicao, M. Ferreira, and P. Steenkiste, "Virtual Traffic Lights in Partial Deployment Scenarios," *Proc. IEEE Intelligent Vehicles Symp.*, June 2013, pp. 988–93.

[15] O. K. Tonguz and W. Viriyasitavat, "Methods and Software for Managing Vehicle Priority in a Self-Organizing Traffic Control System," U.S. Patent Application 14/214,885, 2014.

Biographies

Ozan K. Tonguz (tonguz@ece.cmu.edu) is a tenured full professor in the Electrical and Computer Engineering Department of Carnegie Mellon University (CMU). He currently leads substantial research efforts at CMU in the broad areas of telecommunications and networking. He has published about 300 papers in IEEE journals and conference proceedings in the areas of wireless networking, optical communications, and computer networks. He is the author (with G. Ferrari) of *Ad Hoc Wireless Networks: A Communication-Theoretic Perspective* (Wiley, 2006). In December 2010, he founded the CMU startup known as Virtual Traffic Lights, LLC, which specializes in providing solutions to acute transportation problems using vehicle-to-vehicle and vehicle-to-infrastructure communications paradigms. His current research interests include vehicular networks, wireless networks, sensor networks, self-organizing networks, smart grid, bioinformatics, and security. He currently serves or has served as a consultant or expert for several companies, major law firms, and government agencies in the United States, Europe, and Asia.

Wantanee Viriyasitavat (wantanee.vir@mahidol.ac.th) is a lecturer in the Faculty of Information and Communication Technology at Mahidol University, Bangkok, Thailand, and is also a faculty member in the Department of Telematics, Norwegian Univeristy of Science and Technology, Norway. During 2012–2013, she was a research scientist in the Department of Electrical and Computer Engineering at CMU. She received her B.S./M.S. and Ph.D. degrees in electrical and computer engineering from CMU in 2006 and 2012, respectively. From 2007 to 2012, she was a research assistant at CMU, where she was a member of the General Motors Collaborative Research Laboratory (CRL) working on the design of a routing framework for safety and non-safety applications of vehicular ad hoc wireless networks. She has published more than 30 conference and journal publications, and received numerous awards such as the Dissertation award from the National Research Council of Thailand. Her research interests include traffic mobility modeling, network analysis, and protocol design for ITS.

Results of extensive simulations have shown that, with the proposed scheme, even in a relatively small topology an emergency vehicle is able to arrive at the scene of an accident up to five minutes earlier as compared to a conventional physical traffic lights system. Clearly, for emergency response operations this is quite significant.

High Speed CAN Transmission Scheme Supporting Data Rate of over 100 Mb/s

Suwon Kang, Sungmin Han, Seungik Cho, Donghyuk Jang, Hyuk Choi, and Ji-Woong Choi

A new scheme for enhancing the speed of control area networks has been proposed, where a carrier modulated signal is introduced on top of the existing control area network signal, whereby the data rate can be enhanced over 100 Mb/s

ABSTRACT

As the number of electronic components in the car increases, the requirement for the higher data transmission scheme among them is on the sharp rise. The control area network (CAN) has been widely adopted to support the in-car communications needs but the data rate is far below what other schemes such as Ethernet and optical fibers can offer. A new scheme for enhancing the speed of CANs has been proposed, where a carrier modulated signal is introduced on top of the existing CAN signal, whereby the data rate can be enhanced over 100 Mb/s. The proposed scheme is compatible with the existing CAN network and accordingly enables seamless upgrade of the existing network to support high-speed demand using CAN protocol.

INTRODUCTION

With the exploding interest in connected cars and smart cars, the need for efficient data transmission links within the car is rapidly growing. In past decades, there have been many standards developed to address the requirements for fast and robust links between devices within cars. Table 1 shows a comparison of popular in-car network standards. Among them, the local interconnect network (LIN), control area network (CAN), and FlexRay were developed to support the requirement for connection and control of various in-car devices [1]. LIN is a one-wire communication system developed to meet the need for simple networking of sensor-actuator control requiring low bandwidth.

CAN is the most widely used standard in-car networking technology, featuring simplicity and robust performance. CAN is based on bus topology for simple wiring of individual nodes, and a signal is transmitted and received on a differential pair of unshielded twisted pair (UTP) for robust performance against interference. This robust performance and priority-based fast scheduling has enabled its widespread use in time-critical applications such as engine and power train control. Notwithstanding its dominant adoption across the automotive industry, the current standard of CAN, denoted by CAN 2.0, only supports a maximum data rate of 1 Mb/s [2]. In response to the demand for higher data rates, Bosch recently launched CAN-FD (flexible data rate) supporting data rates of up to 16 Mb/s in an attempt to support the need for higher bandwidth based on widespread adoption of CAN [3]. In addition to increasing the data rate, there have been efforts to analyze and optimize the packet transmission delays in the CAN bus, where many nodes send and receive packets independently [4, 5].

FlexRay was developed to support the requirement for reliable and fast response in applications such as steering, braking, and other safety-critical systems. FlexRay makes use of time slots and cycles instead of random access. Each node is assigned certain time slots in every cycle where it can send packets without any competition or delay. In addition to the above control application, there has been growing demand to connect devices that require much higher bandwidth with the introduction of cameras and entertainment systems in cars.

In order to meet the high data rate demand, automotive Ethernet and media oriented systems transport (MOST) have been developed. Automotive Ethernet is actively being promoted based on its huge success in data networks, both wired and wireless [6]. In order to apply Ethernet to cars, there had to be some modification of standards due to cabling limitations and strict electromagnetic interference (EMI) requirements. As a result, a new Ethernet physical layer standard was developed to support 100 Mb/s over a single pair of copper wires with reduced signaling rates [7]. Unlike data networks, which are less sensitive to packet transmission delay, communication within cars is mostly very time-sensitive. To meet the requirement for real-time communication within cars, the IEEE 802.3br Task Group is working on standards where high-priority packets can interrupt ongoing packet transmission for express delivery [8]. To support in-car infotainment systems, the task of developing efficient multimedia distribution in terms of latency and supported bandwidth is undertaken by the IEEE 802.1 Time Sensitive Networking Task Group. Another Task Group, IEEE 802.3bp, is working to define a standard for 1 Gb/s over a single twisted pair for links up to 15 m, especially for the automotive market [9]. MOST was initially developed to support the need for audio applications [10]. To meet strict EMC requirements, MOST started by using optical fibers as physical media and is based on ring topology to intercon-

Suwon Kang is with GCT Semiconductor Inc.; Sungmin Han, Seungik Cho, and Ji-Woong Choi are with Daegu Gyeongbuk Institute of Science and Technology; Donghyuk Jang is with the Agency for Defense Development, Korea; Hyuk Choi is with the University of Seoul. Ji-Woong Choi is the corresponding author.

0163-6804/16/$25.00 © 2016 IEEE

	LIN	CAN/CAN-FD	FlexRay	Automotive Ethernet	MOST
Physical media	One wire	UTP	Two or four wires	UTP	Optical fiber, coaxial cable, UTP
Multiple access	Master-slave and scheduling	Priority-based messages	Time-division multiple access (TDMA)	Point-to-point link	Priority-based TDMA
Topology	Bus	Bus	Star, bus	Star	Ring
Maximum date rate	19.2 kb/s	1 Mb/s (CAN), 16 Mb/s (CAN-FD)	10 Mb/s	100 Mb/s	150 Mb/s
Target application	Actuator and sensor control such as electronic seat, mirror, and tailgate	ABS, power train, engine control	Steering, braking, and other safety-critical systems.	Infotainment and cameras	Media player and infotainment

Table 1. Comparison of in-car network standards.

nect up to 64 nodes. MOST 25 was first introduced by BMW in 2001, and Toyota introduced MOST 50 in 2007 using unshielded twisted pair (UTP). MOST 150 supports 150 Mb/s using optical cable.

Given that CAN is the most widely deployed of the above standards, this article discusses increasing its data rate significantly to support the requirements for connected cars and smart cars. The following section introduces the CAN standard and protocol. After that, we describe a proposed high-speed CAN signal transmission scheme, and present a transmitter and receiver to implement the scheme. Following that we provide the simulation results for the proposed scheme. Further enhancement of the proposed scheme is then discussed. The final section concludes the article.

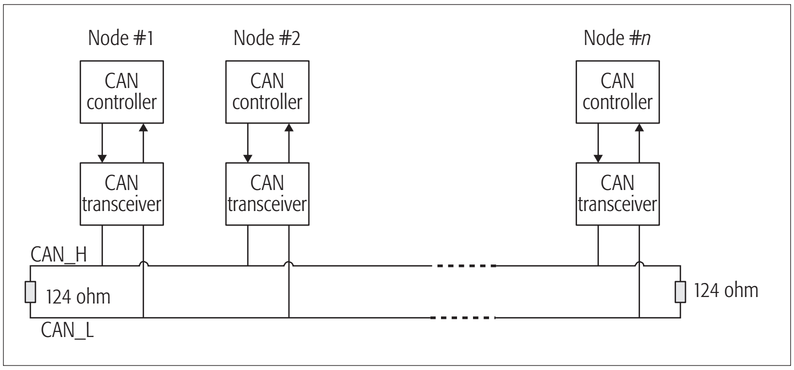


Figure 1. CAN bus configuration.

Controller Area Network System

A conventional CAN system is based on a bus configuration as in Fig. 1. All the nodes share a bus for communication, and each node can transfer data to any node on the bus. Each node comprises a controller and a transceiver. The controller performs bit transmission/reception and timing control. The CAN transceiver is responsible for signal conversion to and from the bus. In the typical application, the two lines on the bus denoted by CAN_H and CAN_L swing from 2.5 V in opposite directions with offset of 1 V, producing a differential signal when transmitting a dominant bit equal to 0. When transmitting a recessive bit equal to 1, the transceiver stops driving the bus, and the levels of the two lines return back to 2.5 V. The resulting CAN signal is transferred on the CAN bus in the form of a differential signal to reduce the interference from other noise sources. The bus is terminated on both ends to prevent signal reflection from degrading the quality of the received signal on each node. On the receiving node, a CAN receiver detects the voltage level difference between CAN_H and CAN_L to determine bits that are being transferred.

Data in the CAN system is transferred in the form of a frame, as shown in Fig. 2. Let the dominant bit and recessive bit be denoted by D and R, respectively. In an idle period, the bus stays at R with no node driving the bus. The transmitting node attempts to send the frame by transmitting D when the bus is confirmed to be in the state of R. The initial transmission of D is called start-of-frame (SOF). Then an arbitration field made up of 12 bits follows SOF. The arbitration field contains a 12-bit field that represents a unique identifier indicating the type of message it wants to send. In the case when two nodes are starting an SOF and transmitting identifiers at the same time, the scheduling is performed based on the priority of the identifiers, using so-called carrier sense multiple access with bitwise arbitration (CSMA/BA). The node with the smaller identifier value prevails and wins the exclusive right to continue to drive the bus. The losing nodes stops transmission and switches to listening state, waiting for the next chance. This unique and fast resolution of contention is one of the greatest benefits that CAN provides compared to other standards such as Ethernet, where conflict of transmission is resolved by each node transmitting after random periods with none of the nodes getting its message transmitted immediately.

The control field contains control bits and bits indicating the number of data following it. In the data field, up to 8 bytes or 64 bits can be transferred in CAN 2.0. In this case, a CAN frame consists of 110 bits for the standard format. The CAN standard specifies that every five consecutive transmissions of the same bits, either Ds or Rs, should be followed by transmission of a different bit, which is inserted at the transmitter, transmitted on the CAN bus, and removed from the receiver. Hence, the actual number of bits transmitted on the CAN bus could be larger than 110 bits due to this dummy bit insertion and removal. In CAN-FD, the duration of the

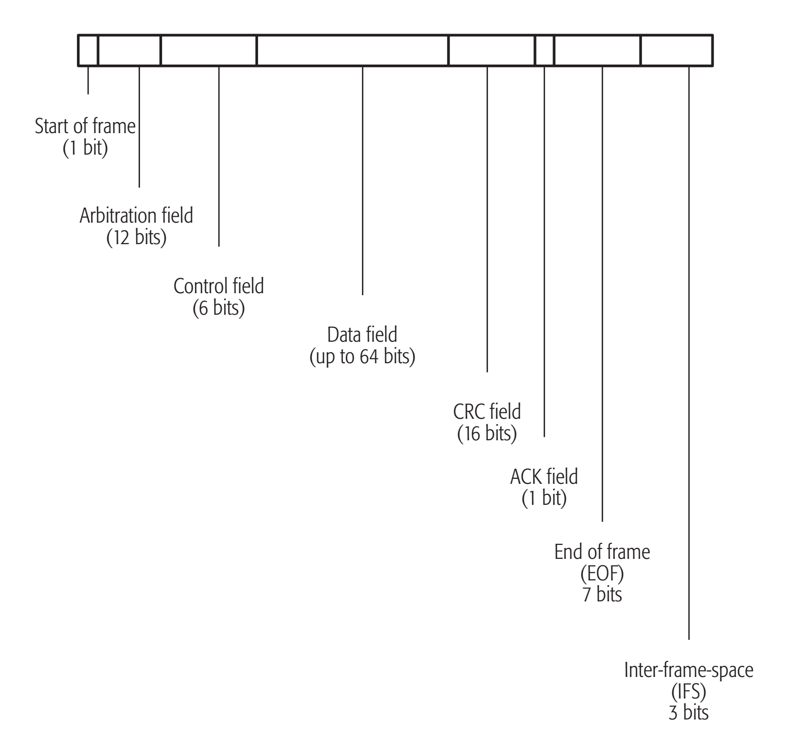


Figure 2. CAN frame.

data field can be increased over 64-bit periods specified in CAN 2.0. The cyclic redundancy check (CRC) field is used to check data integrity. Receiving nodes calculate the CRC of the frame and compare it against the received CRC. The ACK field is used to inform the transmitter of the receipt of the frame. In the case of error, receiving nodes can start transmission of an error frame to report the error. Following ACK are the end-of-frame (EOF) of 7 bits and inter-frame space (IFS) of 3 bits. After IFS, any node can start transmission by sending an SOF.

With many electronic control units (ECUs) and passive nodes connected on the same bus, the priority-based fast contention resolution of the CAN standard is best suited for safety-critical applications. On the flip side, however, nodes with lower-priority messages can suffer from very long delays before getting a chance to transmit, when the bus is mostly occupied by higher-priority nodes. As a solution to this delay issue, a modified CAN standard called time triggered CAN (TTCAN) was proposed and standardized in ISO 11898-4, which is a higher-layer protocol operating on top of a CAN [11]. In TTCAN, there are windows in the time domain during which only one node is allowed to transmit. In addition, there are other windows during which all nodes can compete for transmission. Hence, nodes with lower-priority messages can use windows dedicated to them for data transmission without suffering severe delay.

Another aspect of improvement on the CAN system is related to the improvement of its transmission rate. Most widely adopted CAN systems now support maximum data rates up to 1 Mb/s. As one of the most obvious improvements, overclocking of the CAN signal has been proposed to increase the data rate to over 1 Mb/s [12, 13]. In [12], overclocking of the data field has been proposed, by which the data field in a CAN frame can be transmitted at a higher clock rate, reducing the time spent on data field transmission. However, the overall data rate improvement quickly reaches a certain limit because the maximum of bits in a data field is limited to 64 bits. CAN-FD has been proposed as an upgrade to CAN 2.0 where both overclocking and extension of data field are introduced for higher data rate. The data rate is increased by sending bits in a data field at a higher clocking rate for a longer data field length. However, due to the overclocking during the data field transmission, CAN-FD is not compatible with existing CAN standard devices, which can observe bit transitions within a CAN bit period, subsequently reporting errors. As a solution to avoid this unwanted error reporting during the CAN bit period, partial overclocking within the bit period has been proposed and shown to support higher data transmission up to 16 Mb/s [13]. Even with these improvements, the data rate is still lower than what other standards support to meet the increasing demand for higher data rate.

The Proposed High-Speed CAN System

The primary cause of the data rate limitation of the CAN system comes from three factors. One is the constraint of the bus characteristics, which limits the minimum clock pulse width, which then limits maximum clock rate. Second, due to the attenuation at higher frequency, higher clock pulse suffers from severe edge degradation that could render received waveform hard to detect properly. Finally, only binary signaling is allowed in the standard with very low bandwidth utilization.

The proposed scheme overcomes these limitations by adding carrier modulation to the CAN frame along with higher bandwidth utilization. One of the biggest advantages of using carrier modulation for data transmission is that the proposed system is no longer dependent on edge detection using bit transitions. It also enables the use of higher bandwidth modulation sending multiple bits for each transmit symbol, providing higher data rate without transmission bandwidth increase.

Introduction of the Carrier Modulated Signal

The proposed scheme applies a carrier modulated signal on top of the standard CAN signal when dominant bits are transmitted by a transmitting node. Figure 3a shows the case for the proposed scheme when there are three nodes connected on the bus. The proposed CAN node transmits on the bus while two nodes are receiving from the bus, one being a standard CAN node and the other being a high-speed CAN node. The proposed high-speed CAN transmitter is designed such that the transmitted high-speed CAN signal does not cause any error condition to the existing standard CAN receiver. On the other hand, a high-speed CAN receiver conforming to the proposed scheme can recognize the carrier modulated signal and perform required demodulation. The proposed high-speed CAN controller sends both high-speed CAN transmit bits and standard CAN transmit bits to the transmitter. The standard CAN transmit bits comprise the standard CAN frame as in Fig. 2. High-speed CAN transmit bits are split into two streams for in-phase and quadrature modulation. Depending on the modulation scheme used, the bits in each stream are grouped into $N/2$ bits, where N is an even integer. Thus, for each in-phase/quadrature symbol, N bits are transmitted. The pulse shaping on complex in-phase/quadrature symbols can be additionally applied for band limiting. Then carrier modulation is applied to get the carrier modulated CAN signal, which is applied on top of the standard CAN signal. Then the CAN bus signal generator converts the signal for transmission on the bus.

Figure 3b depicts the high-speed CAN signal generator, where carrier modulated signal $S_p(t)$ is first scaled by A_p and then shifted by fixed offset, typically 1 V, during dominant bit transmission. The resulting high-speed CAN signal $Q_s(t)$ is applied to the CAN bus signal generator only when dominant bits are transmitted in the CAN frame. $Q_s(t)$ is then sent to the bus signal converter to produce differential signal $Q_d(t)$ to be transmitted on the bus. It should be noted that the value of A_p determines the output power of the carrier modulated signal with respect to a standard CAN signal. A higher value of A_p is preferred for reliable transmission of a carrier modulated CAN signal. However, setting A_p too high may cause $Q_s(t)$ to swing below the threshold level of dominant bit detection on the standard CAN receiver. Hence, A_p should be set to a proper value depending on bus configuration and system bit error rate requirements. For instance, A_p could be simply set such that the lowest value of $Q_s(t)$ during a dominant period should be higher than 0.5 V for the purpose of preventing other receivers from erroneously detecting recessive bits. Other sophisticated schemes of A_p can also be considered using the power of $Q_s(t)$.

In an idle period, the proposed high-speed CAN controller starts frame transmission in the same way as the standard CAN controller by sending an SOF and then an arbitration field to gain bus access. If the node wins the bus through the standard arbitration process, high-speed CAN signal transmission can start from the CAN data field. Note that the CAN bits are transmitted as the standard at the maximum rate of 1 Mb/s, while high-speed bits are transmitted on the modulated carrier with higher data rate. When the dominant bit, D, is to be transmitted, the combined signal is the carrier modulated signal shifted by a fixed offset. During recessive bit transmission no signal is transmitted, because the CAN transmitter is allowed to drive the bus only when it is transmitting dominant bits of CAN. When sending R bits, both generation of carrier modulated signal $S_p(t)$ and transmission of high-speed CAN signal stop. It can easily be seen that the period of carrier modulated signal is proportional to the number of dominant bits in the data field of the standard CAN frame. The transmit period of a high-speed CAN signal is maximized when the size the data field is set to the maximum of 64 bits. Further increase in data rate can be achieved when all the data bits are set to dominant 0, enabling the transmission of a high-speed CAN signal over the whole data field, resulting in the highest data throughput. We only consider this case in this article, although it is true that sending meaningful data in the standard CAN frame can bring benefits at the cost of reduction in overall throughput depending on application. It should be noted that the CAN bit stuffing rule dictates putting a bit of opposite polarity after every 5 consecutive bits, while the stuffed bits are removed on the receiver side. Hence, setting data field bits to all Ds would result in a bit pattern of transmission of 5 Ds followed by trans-

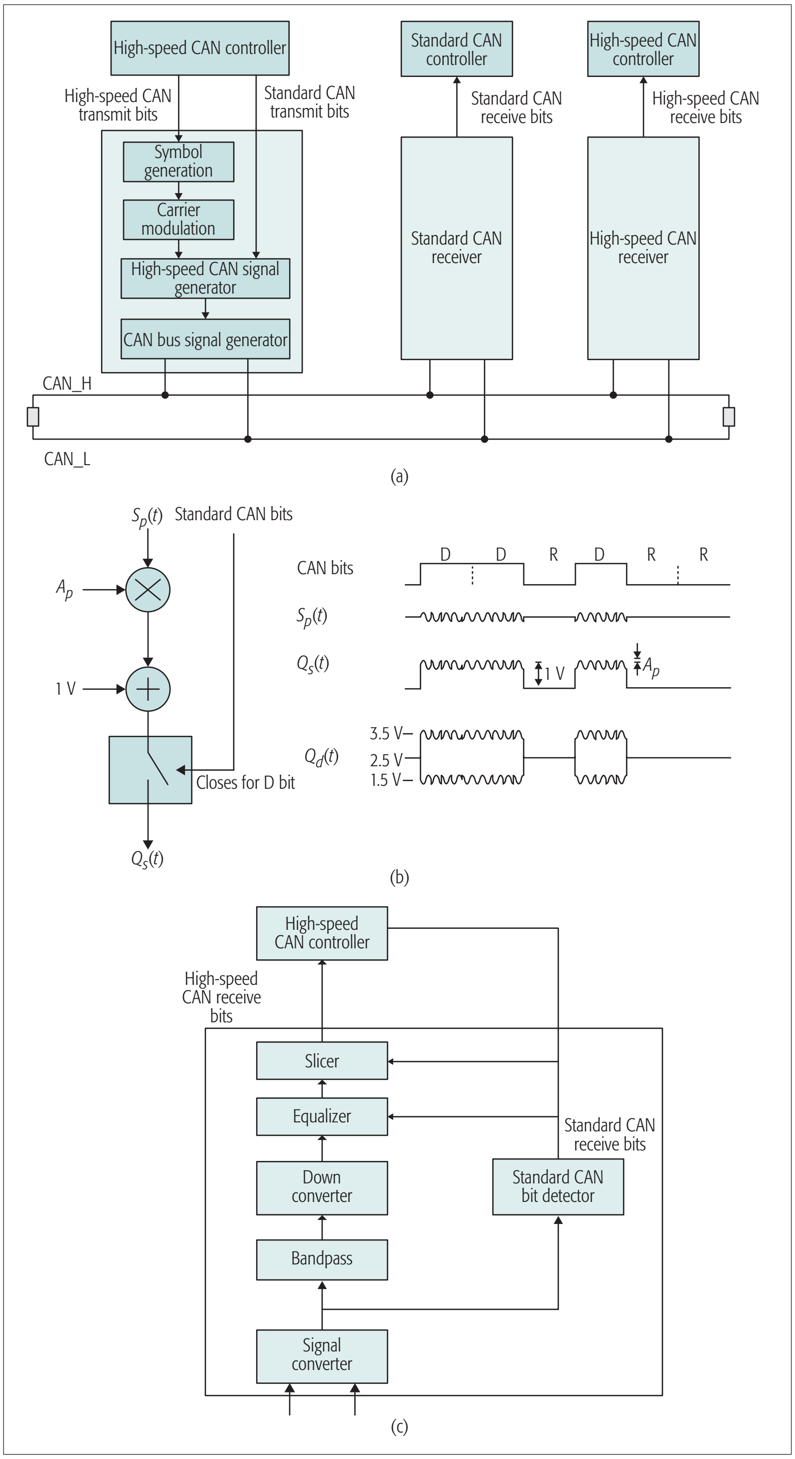


Figure 3. Proposed high-speed CAN system: a) transmitter; b) signal generation; c) receiver.

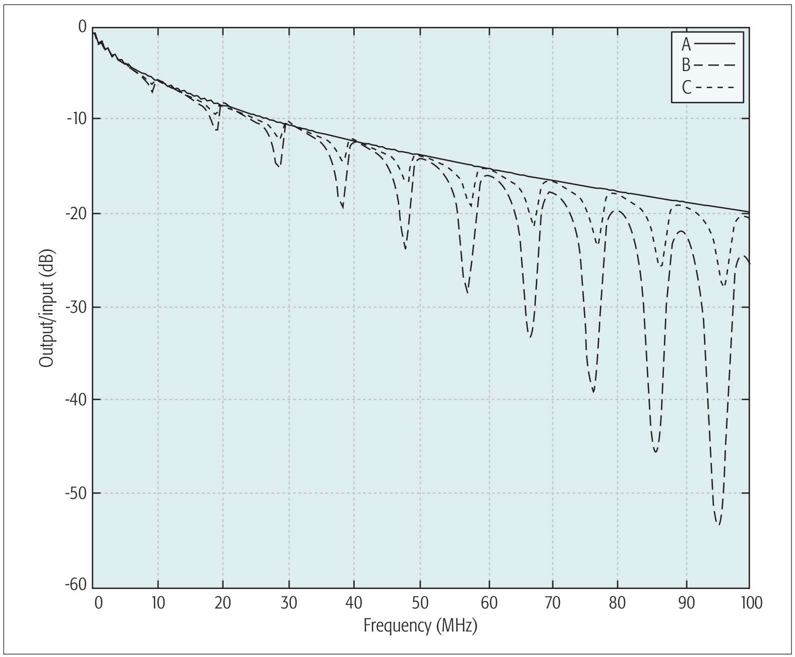


Figure 4. Frequency response of a CAN bus.

mission of one R, which repeats itself during the data field transmission, that is, repetition of the pattern (D-D-D-D-D-R).

Introduction of a Multi-Level Modulation Scheme

In contrast to the binary signaling of the CAN system, the use of multi-level modulation is considered to increase the throughput. High-speed CAN transmit bits are used to form a complex symbol to be modulated and transmitted on the bus. The complex symbol can be constructed using any modulation scheme, including quadrature phase shift keying (QPSK), 16-quadrature amplitude modulation (QAM,) 64-QAM, and so on. Higher order modulation is preferred to get higher data rate. However, the choice of modulation scheme is related to the signal-to-noise ratio, frequency attenuation characteristics of the channel, and receiver complexity.

Receiver

Figure 3c shows the proposed receiver. The differential signal received from the bus is converted to a single-ended signal using a differential-to-single conversion device, which is applied to the standard CAN bit detector and band-pass filter, respectively, for high-speed signal demodulation. The standard CAN detector monitors the bus and detects the start of recessive-to-dominant bit transition during a data field and enables the high-speed CAN demodulator operation to run for a period of 5 D bits. The bandpass filter in the high-speed CAN receiver removes the interference from the standard CAN signal as well as out-of-band noise, providing input to the demodulator and equalizer of the proposed scheme. The down converter removes the carrier to generate a baseband signal providing in-phase and quadrature components. The output of the demodulator is applied to the equalizer for channel compensation. A slicer performs hard decision of equalizer output and generates high-speed CAN received bits.

Simulation Results

Computer simulation is used to verify the performance of the proposed scheme. The simulator models the transmitter and receiver of the proposed scheme, and the output of the transmitter passes through a channel model of the CAN bus before being processed in the receiver.

For ease of explanation, we start with a description of the channel. A CAN bus of 100 m length made of CAT-3 cables is considered, and the transmitter and receiver are placed on each end of the bus with termination. Most of the CAN buses in commercial vehicles are shorter than this, so the length can be considered as the longest case in commercial vehicles. For simplicity, this channel without intermediate nodes is called channel A. In reality, along the bus line, there are many CAN nodes accessing the bus through short feed lines. Starting with the first node 10 m away from the transmitter, a total of 9 CAN nodes are placed 10 m apart, with input impedance of 20 k-ohm. For the length of feed lines, we consider the cases of 0.3 m (channel B) and 0.15 m (channel C) to see their effects on the performance. To obtain the transfer function from transmitter to receiver, we use an ABCD parameter model of transmission lines [15]. The ABCD parameter of each node is found, and the total ABCD parameter from transmitter to receiver is obtained by multiplication of all the ABCD matrices. The transfer function in terms of input voltage to output voltage is obtained and plotted in Fig. 4. Channel A shows monotonically increasing attenuation as frequency increases due to the resistance of the copper lines. Following the overall attenuation trends, channels B and C show additional drops in frequency response due to the reflections from intermediate tap lines. It can be seen that longer tap lines in channel B cause larger drops in the frequency response, rendering equalization in the receiver more challenging. Since there is more attenuation in the high frequency region, it is beneficial to place a high-speed CAN signal in the lower frequency region when high-speed CAN symbol rates are the same. However, placing the signal closer to DC can cause more interference from the standard CAN signal. Hence, trade-offs should be made based on the bus characteristics, carrier frequency, and symbol rate of the high-speed CAN signal. In addition to the distortion from the bus, additive white Gaussian noise (AWGN) is generated and added to channel output.

On the transmitter side, a standard CAN frame with data field bits set to all Ds is generated, and random bit generators are used to generate bits to be carried in a carrier modulated signal. QPSK, 16-QAM, and 64-QAM with normalized rectangular constellations and Gray bit encoding are considered for modulation schemes, and a square root raised cosine filter with roll-off factor of 0.1 is used for the pulse shaping filter. Considering the frequency response of channels in Fig. 4, carrier frequency and symbol rate of high-speed CAN signal $S_p(t)$ are set to 24 MHz and 36 MHz, respectively. The modulated signal is scaled by setting $A_p = 0.3$ and then shifted by 1 V as in Fig. 3b before being applied to the channel input.

On the receiver side, the channel output is applied to the band-pass filter to remove the CAN signal as well as wide-band noise. A linear band-pass filter is designed to have 3 dB bandwidth of 40 MHz centered at carrier frequency with 64 taps. The output is down-converted to baseband, sampled at the symbol rate of the high-speed CAN signal, filtered with the same square root raised cosine filter as the transmitter, and applied to the adaptive equalizer. It is assumed that the timing and carrier synchronization are established for simplicity. An a adaptive decision feedback equalizer is employed to compensate the distortion from the channel. The feed-forward filter and feedback filter employ 24 taps and 8 taps, respectively. For equalizer training, QPSK training symbols generated from random bits are sent for the period of the first 15 D bits out of 64 bits in the data field for each CAN frame. The conventional least mean square (LMS) training method is employed to train equalizer tap coefficients with center tap initialized [14]. Once training is completed, demodulation starts for corresponding data modulation, and the equalizer performs fine adjustment of the equalizer coefficients using slicer output as reference data until the end of the data field. The process repeats itself for each CAN frame.

The bit error rate (BER) performance of the proposed scheme for QPSK, 16-QAM, and 64-QAM are evaluated by varying the signal-to-noise ratio (SNR) of the high-speed CAN signal for channels A, B, and C. SNR is defined to the power of the carrier modulated high-speed CAN signal to the noise power within the bandwidth occupied by the modulated signal. For comparison, a channel with no distortion, denoted by channel R, is considered as a reference. The results are shown in Fig. 5 along with ideal BER performance of each modulation scheme after 100,000 CAN frames are transmitted. Compared to ideal performance, the proposed receiver in the case of channel R shows SNR loss of about 0.9 dB, 1.5 dB, and 1.8 dB for QPSK, 16-QAM, and 64-QAM, respectively, at BER = 10^{-5}. The loss results from non-ideal adaptive equalizer operation and interference from an existing CAN signal. It can be seen that the proposed scheme operates well with reasonable loss in the presence of standard CAN signal interference. Results for channel A show that the loss of the copper lines incurs loss of 3.1 dB, 4.0 dB, and 4.9 dB for QPSK, 16-QAM, and 64-QAM, respectively, at BER = 10^{-5}. It can also be seen that intermediate taps in channels B and C bring about additional loss compared to channel A. This is due to the fact that higher modulation requires higher SNR, and the equalizer cannot fully compensate the larger frequency distortion in channels B and C. The imperfect compensation of distortion results in residual inter-symbol interference (ISI) that degrades SNR at symbol detection. Comparing the results for channels B and C shows higher SNR is required for the case of longer tap lines to achieve the same BER performance. For channel B, the use of QPSK, 16-QAM, and 64-QAM gives loss of about 4.6 dB, 5.9 dB, and 8.0 dB at BER = 10^{-5} compared to ideal performance, while channel C with shorter feed lines shows smaller loss. In a real implementation, the loss can become larger due to impairments such as phase noises of oscillators degrading synchronization performance. To reduce the overall loss, the use of more sophisticated equalizer training needs to be considered to set equalizer coefficients more accurately based on the relatively short period of training symbols. In addition, the use of channel coding such as trellis coded modulation can be considered to lower the required SNR.

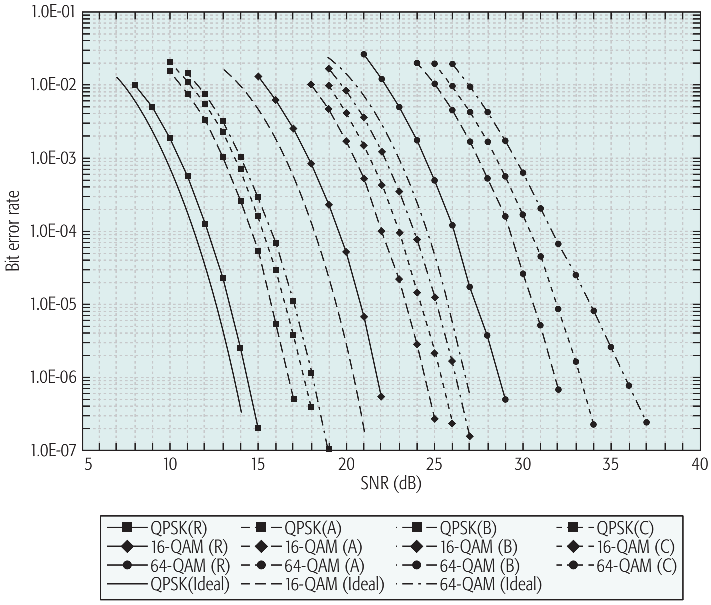


Figure 5. Bit error rate performance of the proposed scheme.

The net data rate of a CAN system is lower than the gross data rate during the data field due to overhead bits in CAN frames such as arbitration field, control field, stuffed recessive bits in the data field, and CRC field. In the CAN standard, the maximum data field is limited to 64 bits, and CAN-FD allows for extension of the data field up to 512 bits. Since the proposed scheme uses the interval of a data field for higher data transmission, net throughput increases as the length of the data field increases. To see the effect of data field length and SNR on the net data rate, throughput from transmitter to receiver in the case of channel B is measured by sending 100,000 CAN frames based on the following assumptions. The transmitter and receiver exclusively make use of the CAN bus, and there is no gap between consecutive frames. Any bit error of the high-speed CAN signal during the data field results in the loss of the entire frame. The receiver counts the number of successfully received bits in received frames and calculates net throughput in terms of bits per second.

Figure 6 shows the net throughput according to the length of the data field and SNR conditions for each modulation scheme. For the proposed scheme, data field lengths of 64 bits and 512 bits are considered. The net throughput of the CAN with a 64-bit data field and CAN-FD with a 512-bit data field is also evaluated for comparison. CAN and CAN-FD show constant throughput of 0.4 and 12.0 Mb/s, respectively, because their robust binary signaling scheme

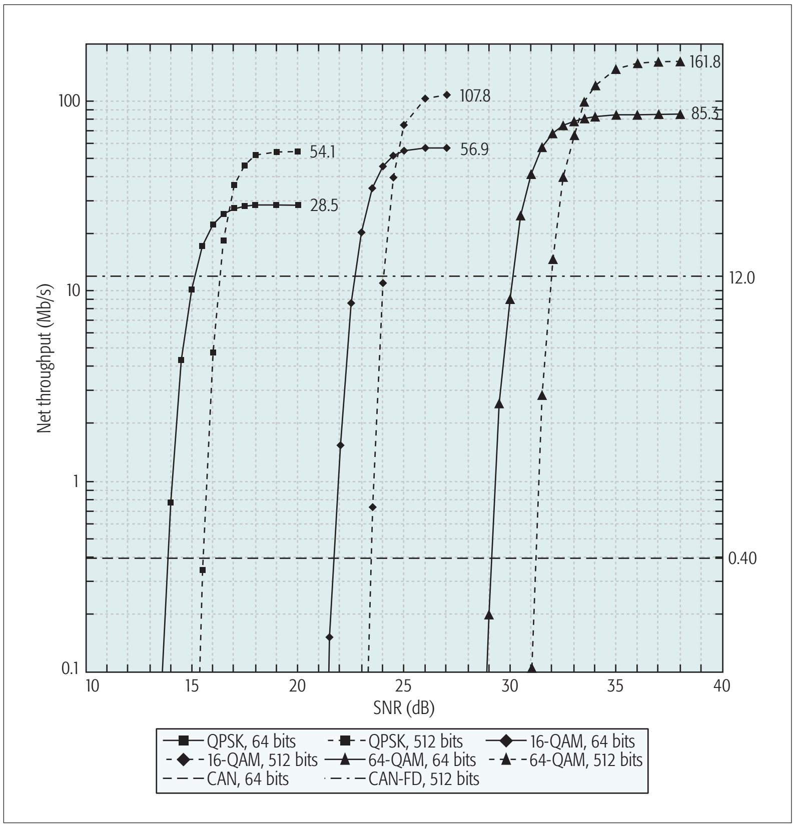


Figure 6. Net throughput of the proposed scheme.

generates no error frames over the range of SNR under consideration. For QPSK, the proposed scheme with data fields of 64 bits and 512 bits shows higher throughput than CAN-FD when SNR is higher than 15.2 dB and 16.3 dB, respectively. Let us define the SNR that gives the same throughput as CAN-FD as cut-off SNR for each modulation scheme for data field length of 512 bits. It can be said that the use of the proposed scheme provides throughput enhancement compared to CAN-FD when the receiver operates above the cut-off SNR for each modulation scheme. We note that 16-QAM and 64-QAM show cut-off SNR of about 24.0 dB and 31.9 dB, respectively. Hence, for example, the use of 16-QAM is recommended when SNR is in the range of 24 to 31.9 dB. It can be seen that the use of a longer data field requires higher SNR to provide the same throughput because a longer frame is more prone to errors than a shorter frame. However, a longer frame can provide higher peak throughput in high SNR conditions. In the current simulation, QPSK with 64 bits and 512 bits provide the maximum throughput of 28.5 and 54.1 Mb/s, respectively, without any bit errors. To provide even higher throughput, 16-QAM and 64-QAM can be employed in higher SNR condition. Note that the maximum net throughput of 107.8 Mb/s and 161.8 Mb/s is achieved for 16-QAM and 64-QAM, respectively, for data field length of 512 bits. If the proposed scheme is to be employed in an existing CAN bus system supporting only 64 bits, the maximum throughput of 56.9 Mb/s and 85.3 Mb/s can be achieved for 16-QAM and 64-QAM, respectively. For each CAN bus in a particular vehicle, the net throughput can be evaluated based on channel characteristics and SNR in order to provide a guideline on the selection of modulation schemes and data field length.

Further Works

While the proposed scheme used simple LMS training with long training sequence, a more efficient equalizer initialization scheme can be used to reduce the length of training bits in the data field to achieve higher throughput, especially for a short CAN frame. Although the frequency and timing of the transmitter and receiver were assumed to be synchronized perfectly for simplicity, the introduction of an appropriate synchronization scheme is required for real implementation. The use of an additional coding scheme can be considered to obtain coding gain and protection against other artifacts such as impulsive noises.

Conclusions

A new scheme for improving the data rate of a CAN system has been proposed by introducing carrier modulation on top of the existing CAN signal. The performance of the scheme in the CAN bus environment has been evaluated in terms of BER and net throughput to show that the proposed scheme can provide higher data rate while keeping backward compatibility with the CAN standard. With the use of a longer frame supported by CAN-FD, the net throughput can be increased to 161.8 Mb/s. The proposed scheme can easily be applied to the existing CAN network without additional deployment of cabling to support the need for high data rate links between devices, resulting in significant reduction of the cost and weight of the vehicle. With the proposed scheme, the CAN standard will be able to expand its application to versatile in-vehicle devices requiring higher data rate beyond its wide adoption to control devices.

Acknowledgments

This work was partly supported by the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korean government (MSIP) (No. B0101-15-0557, Resilient Cyber-Physical Systems Research) and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2015R1A2A2A01008218).

References

[1] S. C. Talbot and S. Ren, "Comparison of FieldBus Systems CAN, TTCAN, FlexRay and LIN in Passenger Vehicles," *Proc. 29th IEEE Int'l. Conf. Distributed Computing Systems Wsps.*, 2009, pp. 26–31.

[2] R. Bosch GmbH, "Controller Area Network (CAN) Specification," v. 2.0., 1991.

[3] R. Bosch GmbH, "CAN with Flexible Data-Rate," v. 1.0, 2012.

[4] K. Tindell and A. Burns, "Guaranteed Message Latencies for Distributed Safety Critical Hard Real-Time Networks," Dept. Comp. Sci., Univ. of York, tech. rep. YCS 229, 1994.

[5] A. Davare, Q. Zhu, and M. D. Natale, "Period Optimization for Hard Real-Time Distributed Automotive Systems," *Proc. 44th IEEE/ACM Design Automation Conf.*, 2007, pp. 278–83.

[6] K. Matheus and T. Königseder, *Automotive Ethernet*, Cambridge Univ. Press, 2014.

[7] K. Matheus, "OPEN Alliance – Stepping Stone to Standardized Automotive Ethernet," *Proc. 2nd Ethernet&IP@Automotive Technology Day*, Regensburg, Germany, 2012.

[8] IEEE 802.3br "Interspersing Express Traffic (IET), 802.3," Nov. 2013.

[9] IEEE 802.3, "Reduced Twisted Pair Gigabit Ethernet Call for Interest," Mar. 2012.

[10] MOST Corp., "MOST150 Inauguration in Audi A3," *MOST Informative*, no. 8, Oct. 2012, pp. 2.

[11] ISO 898-4, "Controller Area Network (CAN) – Part 4: Time-Triggered Communication," 2004.

[12] G. Cena and A. Valenzano, "Overclocking of Controller Area Networks," *Electronics Letters*, vol. 35, no. 22, Oct. 1999, pp. 1923–25.

[13] T. Ziermann, S. Wildermann, and J. Teich, "CAN+: A New Backward-Compatible Controller Area Network (CAN) Protocol with Up to 16x Higher Data Rates," *Proc. Conf. Design, Automation and Test in Europe*, 2009, pp. 1088–93.

[14] D. Jang *et al*., "Communication Channel Modeling of Controller Area Network (CAN)," *Proc. Int'l. Conf. Ubiquitous and Future Networks*, Aug. 2015, pp. 86–88.

[15] Bernard Sklar, *Digital Communications: Fundamentals and Applications*, 2nd ed., Prentice Hall, 2001.

Biographies

Suwon Kang received his B.S., M.S., and Ph.D. degrees in electrical engineering from Seoul National University (SNU), Seoul, Korea, in 1993, 1995, and 2001, respectively. Since 2001, he has been with GCT Semiconductor Inc., California, where he is currently vice president of advanced technology. His research interests are wireless and wire-line communication systems and signal processing.

Sungmin Han received his B.S. degree in electronics engineering from Korea University of Technology and Education, Cheonan, in 2012. Currently, he is working toward his Ph.D. degree in the Department of Information and Communication Engineering (ICE), Daegu Gyeongbuk Institute of Science and Technology (DGIST), Korea. His research areas are communication theory and communication networks.

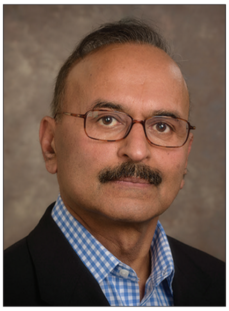
Seungik Cho received his B.S. degree in information and communication engineering from Chungbuk National University (CBNU), Cheongju, Korea, in 2014. Since 2014, he has been studying in the Department of ICE, DGIST, as a Master's student. His research area is advanced communications and RF energy harvesting.

Donghyuk Jang received his B.S. degree in electronics engineering from Kyungpook National University (KNU), Daegu, Korea, in 2013 and his M.S. degree in ICE from DGIST, Korea, in 2015. He is currently a researcher with the Agency for Defense Development (ADD), Korea. His research interests are channel modeling and analysis of communication systems..

Hyuk Choi received his B.S., M.S., and Ph.D. degrees in electrical engineering from SNU in 1994, 1996, and 2002, respectively. In 2003, he joined the faculty at the University of Seoul, where he is currently a professor in the School of Computer Science. His research interests are in information security and signal processing.

Ji-Woong Choi [SM'09] (jwchoi@dgist.ac.kr) received B.S., M.S., and Ph.D. degrees in electrical engineering from SNU in 1998, 2000, and 2004, respectively. From 2005 to 2007, he was at Stanford University, California, as a postdoctoral researcher. From 2007 to 2010, he worked for Marvell Semiconductor, United States, as a staff systems engineer for WiMAX and LTE system design. In 2010, he joined DGIST, where he is currently an associate professor. His research areas are communication theory and signal processing.

RADIO COMMUNICATIONS: COMPONENTS, SYSTEMS, AND NETWORKS



Amitabh Mishra

Tom Alexander

The quality of a mobile channel is severely affected by obstructions such as tall buildings, mountains, and foliage, which in turn limit the performance of wireless communication systems. A transmitted signal propagates toward a receiver generally following standard electromagnetic reflection, diffraction, and scattering mechanisms, which attenuate the strength of the signal, such attenuation being commonly known as propagation loss. There are a number of propagation models with which many of you may be familiar, such as COST 231, Longley-Rice, Okumura, and Hata, to name a few; these have been developed to predict path loss for variable separation distances on the order of hundreds to thousands of meters between transmitters and receivers, and are known as large-scale propagation models. Such models are extensively used in estimating the coverage areas of cellular base stations. Besides the large-scale propagation models, there is a family of small-scale models (also known as fading models) that characterize the signal strength variations for much smaller separation distances, on the order of a few wavelengths to a few tens of meters between transmitters and receivers.

Today's third generation (3G) and 4G cellular and Wi-Fi (IEEE 802.11) carrier frequencies mostly lie between 600–5000 MHz. However, 5G communications systems are envisioned to use 38–60 GHz frequencies, which have wavelengths on the order of millimeters, and are hence denoted as millimeter-wave (mmWave) frequencies. At mmWave frequencies, most objects in the physical environment appear very large relative to the wavelength, thus causing pronounced shadowing losses to the transmitted signal in addition to varying degrees of loss due to reflections, refractions, and diffraction. Because of the smaller wavelengths at mmWave frequencies, the signal is also much more subject to atmospheric attenuation due to rain, snow and hail, oxygen absorption, path depolarization, and other impairments.

A great deal of work is being done today to define what 5G ought be and to identify which technological breakthroughs are fundamental to its successful realization. Driving this revolution are many recently emerged data-hungry applications, such as live streaming video over cellular networks, intra- and inter-vehicular communications, and the IoT (Internet of Things), that require larger bandwidth than what is currently available today. Efforts to shape the 5G landscape are already at work in wireless personal area networks (WPANs, IEEE802.15.3) and wireless local area networks (WLANs, IEEE802.11ad, ah, ax) technology development and standards areas. Due to the widely held view that 5G communications must utilize mmWave frequencies to support the targeted applications, propagation modeling for mmWave communication for large and small separation distances between transmitters and receivers is a very active area of research at the present time in both academia and industry around the globe.

Our first article in this issue of the Radio Communications Series is "Dual Connectivity for LTE Small Cell Evolution: Functionality and Performance Aspects." This article presents the benefits of a user being connected to two base stations at the same time — a macrocell and a small cell — but on two different carriers. In this article, the authors include simulation results showing 50 percent increase in per user throughput and 50 percent reduction in handover failure ratio for mobile users as a result of dual connectivity.

We have also included a second article in this issue, "Radio Propagation Models for 5G Mobile and Wireless Communications," which identifies requirements for 5G radio propagation models depending on use cases and technology trends. This article surveys recent propagation models proposed for mmWave communications, and identifies their strengths and weaknesses. It then presents a map-based propagation model that not only meets the requirements, but also offers extensions to existing stochastic models.

We thank our authors for submitting their cutting edge research contributions to the Radio Communications Series, our reviewers for providing objective and timely reviews to help us select the two articles that are part of this issue, and the readership for their time and attention. In future issues of the Series, we look forward to bringing you similar timely articles from our community of authors covering emerging trends in wireless communications R&D.

Dual Connectivity for LTE Small Cell Evolution: Functionality and Performance Aspects

Claudio Rosa, Klaus Pedersen, Hua Wang, Per-Henrik Michaelsen, Simone Barbera, Esa Malkamäki, Tero Henttonen, and Benoist Sébire

Abstract

DC is one of the most important features introduced in Release 12 of the 3GPP specifications. DC aims at increasing the per-user throughput by improving the utilization of radio resources across two base stations connected via non-ideal backhaul (X2) and operating on different carrier frequencies. By making it possible to maintain the connection to the primary cell located in the macro base station while accessing the extra capacity provided by the small cell layer, DC can also improve the mobility performance in small cell deployments. This article gives an overview of the DC feature as standardized in Release 12. We summarize the supported scenarios and the DC functionality, and also demonstrate by means of detailed system-level simulations how DC can improve end-user throughput and mobility performance.

Introduction

To meet the capacity requirements caused by the continuous increase of mobile broadband traffic, network operators have three main possibilities: bring more spectrum into use, enhance spectral efficiency, and increase cell densification. But with the limited availability of new spectrum and the spectral efficiency of radio access technologies now approaching Shannon's limit [1], increased cell densification through the introduction of small cells is the most promising enabler for increasing the capacity of radio networks. The cost performance ratio of small cells is known to be minimized by having them tightly integrated with the macro layer. Integration of macro and small cells comes in many forms depending on the radio frequency deployment, the type of small cells, and the corresponding inter-node connectivity architecture. Examples of macro and small cell integration options are outlined in [2], corresponding small cell enhancements are discussed in [3], while related multi-cell cooperation techniques are reported in [4]. In general, the tightest integration of small cells is achieved with remote radio heads (RRHs), which utilize joint centralized baseband processing for macro and small cells, assuming high-speed fiber-based fronthaul connections between the nodes. An example of the latter is so-called inter-site carrier aggregation (CA) [4], where a user can be simultaneously connected to a macrocell and a small cell RRH, assuming that those cells use non-overlapping carriers. This form of CA offers promising gains in terms of improved end-user throughput and mobility robustness [4]. The cost of inter-site CA is the need for high-speed, low-latency, fiber-based connections between the RRHs and their corresponding macro site host.

For Third Generation Partnership Project (3GPP) Long Term Evolution (LTE), the macro and small cell integration options are further extended in Release 12 with the introduction of dual connectivity (DC). DC allows users to be simultaneously served by a macro and a small cell operating at different carriers, while the corresponding serving evolved Node-Bs (eNBs) are interconnected with traditional X2-based backhaul connections. These types of backhaul connections are cheaper but characterized by lower capacity and higher latency compared to the high-speed fiber-based fronthaul connections assumed with inter-site CA. Therefore, X2-based backhauls are also referred to as non-ideal backhauls in the remaining of this article. In summary, DC aims to bring some of the benefits of inter-site CA to small cell deployments without fiber-based connection between macro and small cell base stations. The main drawback is represented by the potential impacts on the transport and eNB processing requirements, as we further detail later. The Release 12 DC solution builds on the basics of the CA functionality introduced in Release 10 of the 3GPP specifications [5], as well as on the findings from two earlier LTE Release 12 study items as reported in [6, 7]. The Release 12 DC solution is backward compatible in the sense that a macro and a small cell base station can simultaneously serve Release 12 terminals configured with DC and pre-Release 12 terminals individually without DC.

This article offers a detailed description of the standardized DC feature and related research challenges, focusing on both the overall functionality and performance aspects. This includes describing the deployment scenarios for DC, the functionality and details of the user (U-) and control (C-) plane protocol architectures,

The authors present an overview of the DC feature as standardized in Release 12. They summarize the supported scenarios and the DC functionality, and also demonstrate by means of detailed system-level simulations how DC can improve end-user throughput and mobility performance.

Claudio Rosa, Klaus Pedersen, Per-Henrik Michaelsen, Esa Malkamäki, Tero Henttonen, and Benoist Sébire are with Nokia-Bell Labs; Hua Wang is with Keysight Technologies; Simone Barbera is with Telenor A/S.

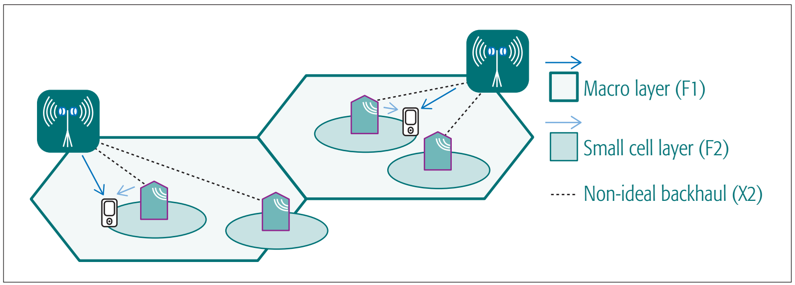


Figure 1. Main deployment scenario of interest for small cell enhancements in general (and DC in particular).

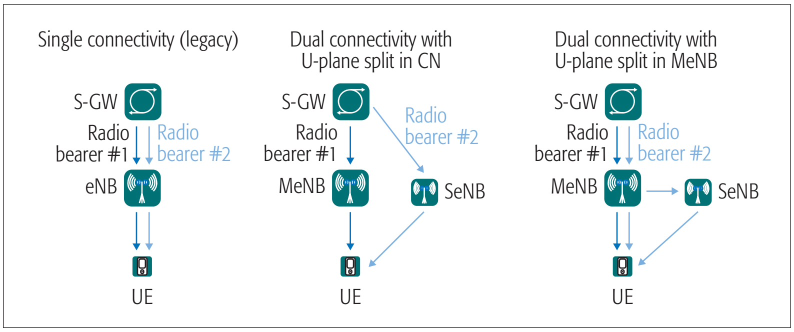


Figure 2. Dual connectivity architecture and U-plane options.

as well as the rationales leading to the various design choices. Mobility and cell management procedures with DC are outlined as well. A set of system-level performance results is presented to demonstrate the benefits of DC. Throughout the article, guidance on how to best use and operate the many options that come with the introduction of DC are presented. Finally, concluding remarks and an outlook toward future research and DC enhancements in upcoming 3GPP releases are presented.

Deployment Scenario

A deployment scenario for DC is depicted in Fig. 1. It corresponds to a dedicated carrier small cell deployment where different carrier frequencies are used by the macro (F1) and small cells (F2), the macro and small cells being interconnected via non-ideal X2 backhauls. Although this is the typical deployment scenario for DC, the use of DC is not necessarily restricted between a macro and a small cell eNB, but can in principle be applied between any pair of eNBs that are interconnected via X2 and operate on at least two different carrier frequencies. DC is designed to work for backhaul latencies of up to 60 ms [8], although its applicability in scenarios with larger latencies is not precluded.

DC deployment scenarios cover both time-synchronized and time-asynchronous networks, and cases with frequency- and time-division duplex (FDD and TDD), and also allow the use of CA at each layer. Multiple carriers can be deployed at both macro and small cell layers, although the aggregated bandwidth per user equipment (UE) cannot exceed either 100 MHz or 5 component carriers as per Release 10 CA assumptions [5]. Without loss of generality, in this article we focus on outdoor small cell deployments with synchronized operation of macro and small cell base stations, assuming FDD for both macro and small cell layers, and only two component carriers configured (one in the macro layer and one in the small cell layer).

Dual Connectivity Solution

When configured with DC a UE is simultaneously connected to two eNBs: a master eNB (MeNB) and a secondary eNB (SeNB). The MeNB and SeNB are connected via an X2 interface. In LTE, UEs can be in either of two modes: radio resource control (RRC) connected mode and idle mode. When in RRC connected mode, the UE has a radio connection with the network, and can transmit and receive U-plane data. When in idle mode, the UE can only be paged and/or initiate communication with the network via a random access procedure. In any case, an idle mode UE needs to initiate a radio connection with the network and switch to RRC connected mode before it can transmit and receive U-plane data. DC is only applicable to UEs in RRC connected mode. The U-plane and C-plane alternatives to support DC are summarized in the remaining part of this section.

User Plane

From a user plane perspective, 3GPP has standardized two types of DC solutions: one with split of U-plane data in the core network (CN), and another one with split of U-plane data in the MeNB (Fig. 2). With data split in the CN, each eNB involved in DC has its own user plane connection toward the serving gateway (S-GW). Thus, the MeNB and SeNB serve separate radio bearers. In other words, when a radio bearer is configured to carry data from one or several applications, the corresponding data can only be transmitted from and toward one of the eNBs involved in the DC configuration (unless a radio bearer reconfiguration is performed). Quite differently, with data split in the MeNB, only the MeNB has a user plane connection toward the S-GW, and data from one radio bearer can be transmitted via both the MeNB and SeNB.

Advantages and Drawbacks of DC Solutions: The main advantage of splitting U-plane data in the MeNB is the higher flexibility in simultaneously using the spectrum available in the macro and small cell layers. However, this typically comes at the cost of increased transport and processing capacity in the MeNB since all data transmitted toward the UE always has to traverse the MeNB. Moreover, due to X2 latency, such option requires distributed radio resource management (RRM), as well as flow control of U-plane data between the MeNBs and the SeNBs. Some of these challenges are better exemplified in the remainder of this section, while the performance impacts of distributed RRM and X2 latency are illustrated below. In contrast, splitting user plane data in the CN can achieve some of the offloading gains, but cannot fully exploit the peak data rate and fast inter-layer load balancing gains offered by DC with U-plane data split in the MeNB. This is due to the limit imposed by the one-to-one mapping between radio bearers and eNBs involved in the DC configuration.

The two DC solutions are in practice realized by the standardization of three different types of radio bearers: radio bearers that are served by the MeNB alone, called master cell group (MCG) bearers; radio bearers that are served by the SeNB alone, called secondary cell group (SCG) bearers; and radio bearers that are served by both the MeNB and SeNB, called split bearers. Since the two DC solutions require different transport and processing capabilities, the Rel-12 specifications do not support simultaneous configuration of SCG and split bearers for the same UE. Figure 3 illustrates the user plane protocol stack at the MeNB, SeNB and at the UE in case the UE is configured with one MCG and one SCG bearer (Fig. 3a), or one MCG and one split bearer (Figs. 3b and 3c). The solid and dashed arrows illustrate the flow of U-plane data in downlink and uplink, respectively. It can be observed from the two rightmost configurations of Fig. 3 that Release 12 specifications do not support splitting U-plane data in the uplink, in which case only routing is supported; that is, for a split bearer the UE is configured in terms of whether it should transmit U-plane data in uplink toward the MeNB (Fig. 3b) or toward the SeNB (Fig. 3c).

Apart from that, the DC solutions are designed to maintain as much as possible the general structure of the LTE link layer design [9]. The packet data convergence protocol (PDCP) is mainly responsible for ciphering (security) and header compression. The main functions provided by the radio link control (RLC) layer are segmentation, concatenation, and in-order delivery of packet data units to higher layers. The medium access control (MAC) is essentially responsible for scheduling, multiplexing, and retransmission of packet data, while the physical layer (PHY) is primarily accountable for transmitting the packet data over the air interface.

Upper Layer 2 Protocol Stack: With the split bearer option of DC (Figs. 3b and 3c), the PDCP functionality in the UE needs to be updated to be able to perform reordering of packet data units delivered via two independent RLC entities. This is done using a reordering window controlled by a reordering timer running at the UE. The reordering function of PDCP is designed to work within the backhaul latency requirements specified in [8].

Lower Layer 2 Protocol Stack: Two separate MAC entities (one per eNB) are introduced for UEs configured with DC: one handling the scheduling of transmissions in the MGC and one handling the scheduling of transmissions in the SGC. In LTE, discontinuous reception (DRX) allows the UE to decrease its power consumption in connected mode. DRX operation is tightly coupled to scheduling decisions, and with two MAC entities corresponding to two schedulers separately located in the MeNB and SeNB, a common DRX is not possible. Although some coordination between the eNBs for DRX configuration is seen as beneficial from the UE power consumption perspective, how to achieve the coordination is not specified but rather left for network implementation.

Buffer Status Reporting and Power Headroom Reporting: Buffer status reporting and power headroom reporting were introduced in LTE Release 8 with the scope to assist and control scheduling of uplink data transmissions in the eNB. With dual connectivity, power headroom reports transmitted toward one eNB include power headroom information on all the activated cells in the UE, that is, the activated cells served by the other eNB involved in the DC configuration as well. For non-split bearers, buffer status reports of MCG bearers are transmitted toward the MeNB, while buffer status reports of SCG bearers are transmitted toward the SeNB. For split bearers, the data in the PDCP buffer is only considered for buffer status reporting toward the eNB configured for U-plane data transmission in the UL.

In order to clarify the UE behavior, let us consider a UE configured with DC and having one cell configured in the MeNB and one cell configured in the SeNB. Moreover, the UE is configured with one MCG and one split bearer. When reporting power headroom information, the UE considers the allocated transmission power on each cell and compares it to the maximum allowed transmission power. The difference between the maximum and the allocated power is the power headroom. The UE reports the power headroom measured on all activated cells to both the MeNB and SeNB. The UE also reports the amount of data belonging to the MCG bearer to the MeNB, while the amount of data belonging to the split bearer is reported to either the SeNB or the MeNB, depending on toward which node the UE is configured to transmit UL user plane data.

Random Access: In order to maintain the PHY connection to both eNBs involved in DC, independent (and parallel) random access procedures are supported in the MeNB and SeNB. For example, a UE that has lost synchronization to the SeNB (e.g., due to expiration of the advanced timer) can initiate a random access

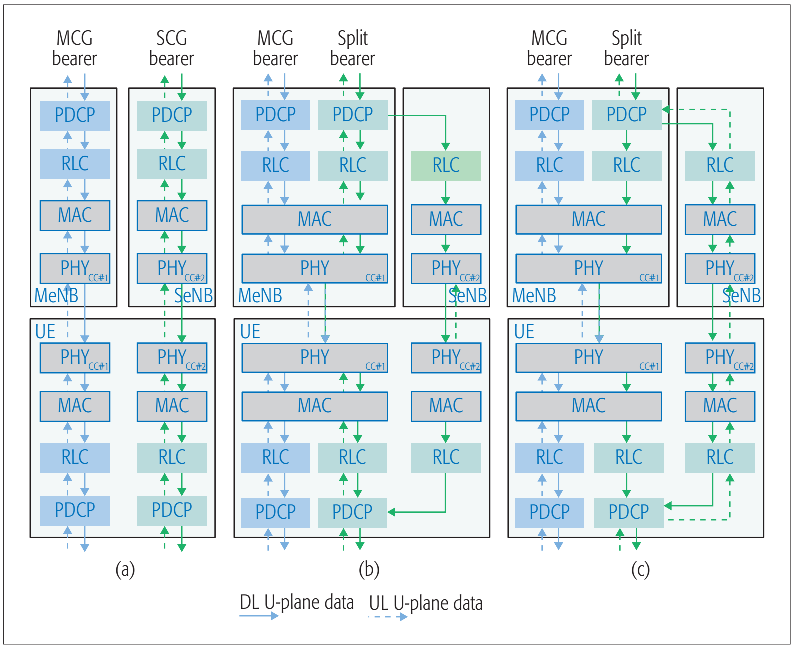


Figure 3. Dual connectivity DL/UL U-plane protocol stacks at eNB and UE with a few example configurations: a) MCG bearer + SCG bearer + split bearer, with UL configured in either b) MCG or c) SCG.

In line with the basic principles of LTE, network-controlled UE-assisted mobility and cell management also applies to DC. Since the RRC always resides in the MeNB, the MeNB also maintains the UE RRM measurement configuration and acts according to the received measurement reports.

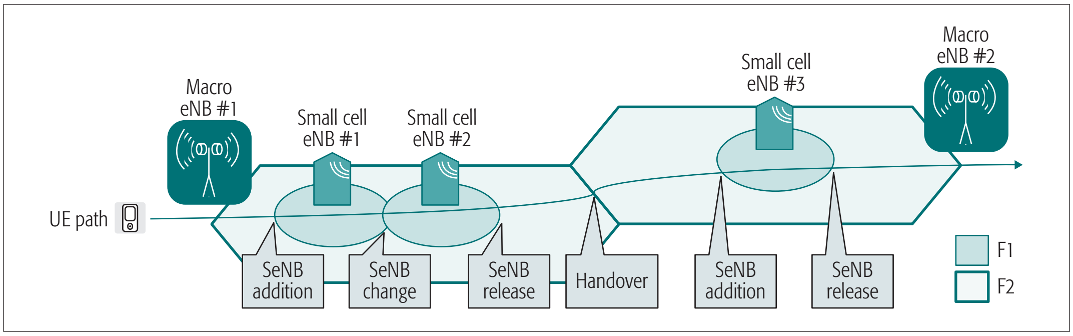


Figure 4. UE moving in macrocell coverage area with several small cells.

procedure toward the SeNB without affecting data transmission and reception with the MeNB.

Uplink Power Control: With simultaneous UL transmissions toward multiple eNBs in DC, UL power control becomes more challenging. The possibility of semi-statically configuring separate maximum transmission powers for UL transmission toward the MeNB and SeNB is therefore introduced. In this way, a minimum amount of power can be guaranteed for UL transmissions toward one eNB even if (due to uncoordinated scheduling) the other eNB is already trying to fully utilize the maximum UE power capabilities. The power control mechanism for DC still allows using the full transmission power for UL transmission toward one eNB if there are no simultaneous UL transmissions toward the other eNB.

X2 Enhancements: DC also imposes enhancements for the X2 specifications [10]. First, each PDCP transmitting entity in DL needs to get from lower layers regular indication of successful delivery of PDCP packets. In LTE this service is provided by the RLC layer. However, in the case of split bearers, the RLC entity responsible for reliable data transmission via the SeNB and the PDCP entity are located in separate eNBs (Figs. 3b and 3c). Therefore, with a split bearer the SeNB regularly provides the MeNB (over X2) with information on the last in-sequence successfully delivered PDCP packet among the ones it has received from the MeNB. Also, the MeNB needs to decide which data is to be transmitted via its own cells, and which data is to be forwarded to the SeNB over X2. Since the MeNB does not necessarily know the conditions (channel, interference, load, etc.) in the SeNB, X2 flow control is necessary in order to avoid data overflow and underflow in the SeNB. Therefore, the X2 specifications are enhanced, enabling the SeNB to send radio bearer level information on its available buffer size to the MeNB. This standardized signaling can be used to implement request-based flow control mechanisms such as the one presented in [11], which for X2 latencies up to 20 ms is shown to provide acceptable performance with respect to small cell deployments, which require high-speed and low-latency interfaces with the macro eNB.

Control Plane

When a UE is configured with DC, the MeNB maintains the RRC connection, and the control plane connection toward the mobility management entity (MME) is always terminated in the MeNB. This means that there is no impact of DC on the MME and on the CN in the case of a split bearer, and only minor impact in the case of an SCG bearer due to U-plane and C-plane connections of the same UE terminating at different eNBs. The MeNB controls the DC configuration: it generates and sends all the RRC messages to the UE. Transmission of RRC messages via the SeNB is not supported. However, the SeNB can request the MeNB to change or release its own part of the RRC configuration via an X2 message.

Radio link monitoring (RLM) is one important procedure in LTE. It is used to monitor the radio link conditions so that appropriate actions can be taken if a radio link failure (RLF) occurs. With DC, a UE only performs RLM on two cells: one in the MeNB (the primary cell, PCell) and one in the SeNB (the primary secondary cell, PSCell). The difference between RLM in the MeNB and SeNB is that upon detection of an RLF in the SeNB, the UE does not trigger the RRC connection re-establishment procedure because the RRC connection toward the MeNB can still be maintained even if the radio link to the SeNB fails.

Mobility and Cell Management

In line with the basic principles of LTE, network-controlled UE-assisted mobility and cell management also applies to DC. Since the RRC always resides in the MeNB, the MeNB also maintains the UE RRM measurement configuration and acts according to the received measurement reports. As in CA, the UE can be configured to perform measurements from its serving and surrounding cells. For additional details and background information on the RRM measurement events for mobility and secondary cell (SCell) management, see [12, 13]. RRM measurements with DC are largely unaffected compared to CA, except for modifications to RRM measurement events A3 (neighbor becomes offset better than PCell) and A5 (PCell becomes worse than threshold #1 and neighbor becomes better than threshold #2). With DC, the PSCell can also be used instead of the PCell for those events.

To exemplify the procedures involved in DC operation, Fig. 4 shows a UE configured with DC moving along the picture trajectory (UE path). When the UE enters the coverage area of small cell eNB #1, the SeNB addition procedure is first used to configure the UE with DC. Based on RRM measurements reported by the UE, the MeNB can then decide to initiate the SeNB

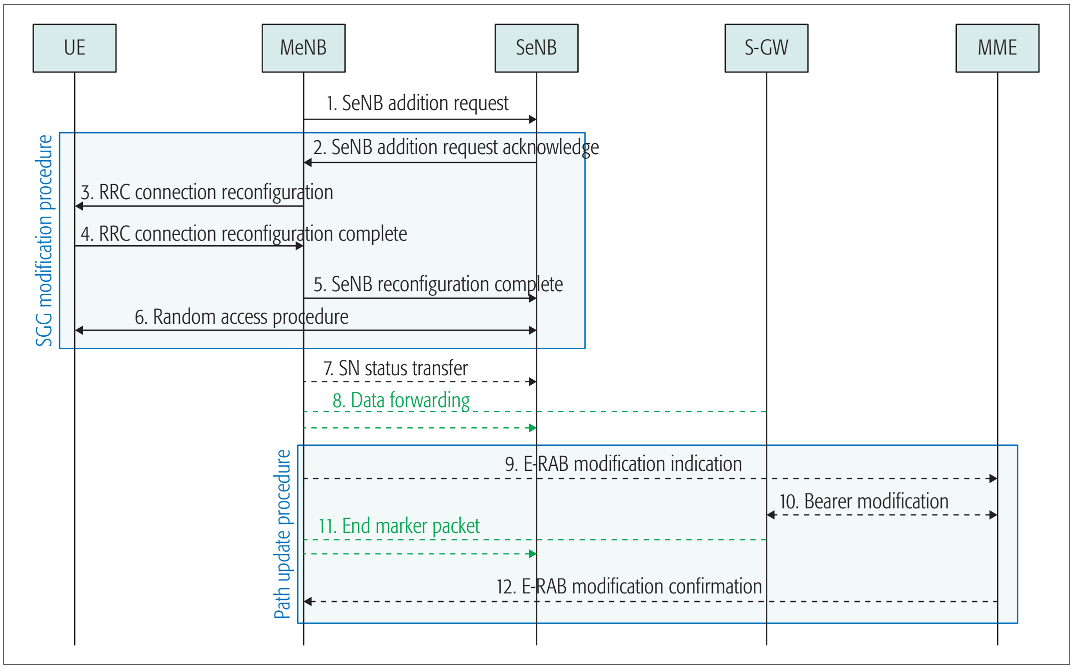

UE
MeNB
SeNB
S-GW
MME
SGG modification procedure
1. SeNB addition request
2. SeNB addition request acknowledge
3. RRC connection reconfiguration
4. RRC connection reconfiguration complete
5. SeNB reconfiguration complete
6. Random access procedure
7. SN status transfer
8. Data forwarding
Path update procedure
9. E-RAB modification indication
10. Bearer modification
11. End marker packet
12. E-RAB modification confirmation

Figure 5. SeNB addition procedure.

If the radio link to the SeNB deteriorates due to the UE moving out of the corresponding small cell, the MeNB can finally decide to initiate the SeNB release procedure. Upon or before doing a handover the SeNB must be (temporarily) released. If the data connection with the previously configured SeNB (or alternatively with another small cell eNB) is still possible after the handover, DC can be restored using the SeNB addition procedure.

change procedure to change the SeNB for the UE from small cell eNB #1 to small cell eNB #2, involving the release of the current SeNB and the addition of a new SeNB. If the radio link to the SeNB deteriorates due to the UE moving out of the corresponding small cell, the MeNB can finally decide to initiate the SeNB release procedure. Upon or before doing a handover the SeNB must be (temporarily) released. If the data connection with the previously configured SeNB (or alternatively with another small cell eNB) is still possible after the handover, DC can be restored using the SeNB addition procedure.

As an example of the signaling involved with the DC (re)configuration procedures, Fig. 5 shows the signaling flow chart for the SeNB addition procedure as follows:

- Upon requesting the SeNB to allocate radio resources for a specific UE/radio bearer (1), the MeNB also indicates to the SeNB the bearer characteristics. In addition, the MeNB indicates to the SeNB the configuration of the cells under MeNB control and the UE capabilities, from which the SeNB can infer the amount of radio resources it is allowed to use. The MeNB can also provide the latest measurement reports for the cell(s) requested to be added in the SeNB.
- The SeNB may reject the request. However, if the RRM entity in the SeNB is able to admit the resource request, it allocates respective radio and transport network resources. The SeNB provides the new radio resource configuration of the cells under its control to the MeNB (2).
- The MeNB then sends an RRC reconfiguration message to the UE (3) including the new radio resource configuration of the cells under control of the SeNB.
- The UE applies the new configuration and replies to the MeNB with an RRC reconfiguration complete message (4).
- The MeNB informs the SeNB that the UE has completed the reconfiguration procedure successfully (5). The UE performs synchronization with the primary cell of the SeNB, that is, the PSCell (6). Note that the random access procedure toward the SeNB and the indication of RRC reconfiguration completion/failure toward the MeNB are performed independently.
- The MeNB may take action to minimize service interruption due to activation of DC by initiating data forwarding toward the SeNB (7, 8).
- Finally, for SCG bearers only, the user plane path toward the CN also needs to be updated (9–12).

Performance Results

The performance benefits of using DC with a split bearer are illustrated in the following for 3GPP Release 12 scenario 2a as defined in [7], by means of extensive system-level simulations. The results also show the performance of DC vs. that of intersite CA with high-speed fiber-based fronthaul connections between an MeNB and small cell nodes. The network topology consists of a standard hexagonal grid of three-sector MeNBs complemented by a set of outdoor small cells. The small cells are randomly deployed in condensed clusters with four small cells per macro sector area. The simulator follows the LTE specifications, including detailed modeling of major RRM functionalities. A dynamic birth-death traffic model is applied to generate user calls, where call arrival is according to a Poisson process with arrival rate per macro cell area. Each call has a finite payload size of B = 4 Mb. Thus, the average offered load per macrocell area equals $\lambda \times B$ Mb/s. Channel-aware joint proportional fair scheduling is assumed [14]. The X2 flow control algorithm is a simple request-and-forward-based scheme, which aims to keep the transmission buffer depth in the SeNB to a pre-

Due to the possibility of performing radio aggregation across two eNBs, the user throughput performance with DC is significantly higher than without DC, but still relatively close to the performance of inter-site CA with fiber-based fronthaul connections between macro and small cell eNBs.

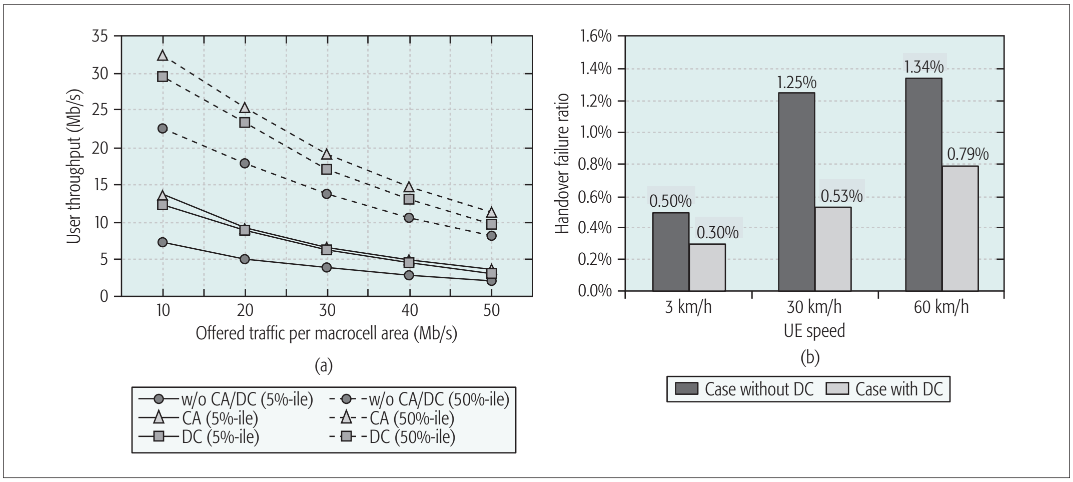


Figure 6. Downlink user a) throughput; b) mobility performance with and without DC.

defined target level [11]. Requests from the SeNB to the MeNB are sent periodically every 5 ms on a per-user basis. Received data from the MeNB are buffered in the SeNB until they have been successfully transmitted to the UE. The X2 latency is assumed to equal 5 ms. More detailed system level simulation parameters can be found in [11].

Figure 6a shows the 5th percentile and 50th percentile end-user throughput as a function of the offered traffic per equivalent macrocell area. The 5th and 50th percentile user throughput denote the throughput experienced by at least 95 and 50 percent of the end users, respectively. Due to the possibility of performing radio aggregation across two eNBs, the user throughput performance with DC is significantly higher than without DC, but still relatively close to the performance of inter-site CA with fiber-based fronthaul connections between macro and small cell eNBs. If having a target to serve, for example, 95 percent of the users with at least 4 Mb/s, it is observed from Fig. 6a that the maximum tolerable offered load increases from 30 Mb/s (without DC) to approximately 45 Mb/s for cases with DC, corresponding to a capacity gain of about 50 percent. At the same time, the performance degradation of DC compared to inter-site CA is almost negligible. Additional DC performance results can be found in [11, 14].

DC also offers improved mobility robustness. Figure 6b shows the handover failure (HOF) ratio for cases with and without DC. In this context, a HOF event is declared if a radio link failure for the MeNB occurs after the time to trigger expires, which means during the handover preparation or execution time, as defined in [15]. A radio link failure is declared when the downlink signal-to-interference-plus-noise ratio (SINR) for the UE on the primary cell has been below –8 dB and stayed below –6 dB for the duration of a predefined timer, here assumed to be 1 s. It can be generally observed that the HOF percentage is significantly lower with DC enabled, thus demonstrating that the use of DC also offers benefits in terms of mobility robustness. The improved HOF performance from using DC comes from always having the PCell at the macro layer, while utilizing the small cells whenever possible for the user. Keeping the PCell at the macro layer essentially means that the HOF probability corresponds to the macro-only scenario, and therefore is not affected by the small cells.

Concluding Remarks

In this article we have described the DC concept for LTE-Advanced. The DC feature brings carrier aggregation gains to small cell deployments with non-ideal X2 backhaul connection. System-level simulation results carried out in representative small cell deployment scenarios show a capacity increase from DC with a split bearer on the order of 50 percent, while 95 percent of users can still experience user throughput of at least 4 Mb/s. At the same time, the mobility robustness performance can be improved by significantly reducing the handover failure ratio due to the UE always being connected to the macro layer even when utilizing the radio resources in the small cell layer. Further evolution of DC includes improvement of the mobility and small cell management procedures (handover while retaining the DC configuration and handover with SeNB addition), as well as support of uplink bearer split and local breakout solutions for DC. These enhancements have already been included in Release 13 of the LTE specifications. Future research activities should target reducing the impact on the transport and processing requirements of DC, while maintaining the advantages of radio aggregation across eNBs. Also, future research may consider the applicability of DC to cloud network architectures, as well as the extension of DC to inter-radio-access-technology deployments such as DC between LTE and 5G.

References

[1] C. E. Shannon, "Communication in the Presence of Noise," *Proc. Inst. Radio Engineers*, vol. 37, no. 1, 1949, pp. 10–21.

[2] Y. Kishiyama *et al.*, "Future Steps on LTE-A: Evolution toward Integration of Local Area and Wide Area Systems," *IEEE Wireless Commun.*, vol. 20, no. 1, Feb. 2013, pp. 12–18.

[3] T. Nakamura *et al.*, "Trends in Small Cell Enhancements in LTE-Advanced," *IEEE Commun. Mag.*, vol. 51, no. 2, Feb. 2013, pp. 99–105.

[4] B. Soret *et al.*, "Multicell Cooperation for LTE-Advanced Heterogeneous Network Scenarios," *IEEE Wireless Commun.*, vol. 20, no. 1, Feb. 2013, pp. 27–34.

[5] K. I. Pedersen *et al.*, "Carrier Aggregation for LTE-Advanced: Functionality and Performance Aspects," *IEEE Commun. Mag.*, vol. 46, no. 6, June 2011, pp. 89–95.

[6] 3GPP TR36.842, "Study on Small Cell Enhancements for E-UTRA and E-UTRAN – Higher Layer Aspects," v. 12.0.0, Dec. 2013.

[7] 3GPP TR36.872, "Small Cell Enhancements for E-UTRA and E-UTRAN – Physical Layer Aspects," v. 12.1.0, Dec. 2013.

[8] 3GPP TR36.932, "Scenarios and Requirements for Small Cell Enhancements for E-UTRA and E-UTRAN," v. 12.1.0, Dec. 2013.

[9] A. Larmo *et al*., "The LTE Link-Layer Design," *IEEE Commun. Mag*., vol. 47, no. 4, Apr. 2009, pp. 52–59.

[10] 3GPP TS36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2 (Release 12)," v. 12.4.0, Dec. 2014.

[11] H. Wang *et al*., "Inter-eNB Flow Control for Heterogeneous Networks with Dual Connectivity," *Proc. 2015 IEEE 81st VTC*, Glasgow, U.K., 11–14 May 2015, pp. 1–5.

[12] K. I. Pedersen *et al*., "Mobility Enhancements for LTE-Advanced Multilayer Networks with Inter-Site Carrier Aggregation," *IEEE Commun. Mag.*, vol. 51, no. 5, May 2013, pp. 64–71.

[13] S. Barbera *et al*., "Mobility Analysis for Inter-Site Carrier Aggregation in LTE Heterogeneous Networks," *Proc. 2013 IEEE 78th VTC*, Las Vegas, NV, 2–5 Sept. 2013, pp. 1–5.

[14] H. Wang *et al*., "Dedicated Carrier Deployment in Heterogeneous Networks with Inter-Site Carrier Aggregation," *Proc. IEEE WCNC*, Shanghai, China, 7–10 Apr. 2013, pp. 756–60.

[15] 3GPP TR36.839, "Evolved Universal Terrestrial Radio Access (E-UTRA); Mobility Enhancements in Heterogeneous Networks," v. 11.1.0, Jan. 2013.

Biographies

Claudio Rosa (claudio.rosa@nokia.com) received his M.Sc.E.E. and Ph.D. degrees in 2000 and 2005, respectively, from Aalborg University. In 2003 he also received an M.Sc.E.E. degree in telecommunication engineering from Politecnico di Milano, Italy. He is currently with Nokia-Bell Labs, Aalborg, Denmark, where he works as a senior wireless network specialist. At present his research activities are mainly focused on 3GPP standardization of LTE-Advanced, more specifically on small cell enhancements and unlicensed spectrum opportunities for cellular deployments.

Klaus I. Pedersen (klaus.pedersen@nokia.com) received his M.Sc. degree in electrical engineering and Ph.D. degree from Aalborg University in 1996 and 2000, respectively. He is currently a senior wireless network specialist at Nokia-Bell Labs, and a part-time professor at Aalborg University in the Wireless Communications Network (WCN) section. He is the author/co-author of more than 100 peer-reviewed publications on a wide range of topics, as well as an inventor on several patents. His current work is related to 5G air interface design, including radio resource management aspects and system-level performance assessment. He is also currently contributing to the EU funded research project FANTASTIC-5G focused on development of a new multi-service 5G air interface for below 6 GHz operation.

Hua Wang received his Bachelor's degree from Zhejiang University, China, in 2003, and his M.Sc. and Ph.D. degrees from the Technical University of Denmark, Copenhagen, in 2005 and 2009, respectively. From 2008 to 2009 he was a visiting scholar at the University of Texas at Austin. In 2009 he joined Aalborg University as a postdoctoral researcher and became an associate professor in 2013. In 2016 he joined Keysight Technologies, Aalborg, as a senior researcher. His main research interests are in the area of radio resource management and performance analysis for 4G/5G systems.

Per-Henrik Michaelsen (per_henrik.michaelsen@nokia.com) received his M.Sc. degree in electrical engineering from Aalborg University in 1982. He is currently a wireless networks specialist at Nokia-Bell Labs. He is the author/co-author of more than 25 peer-reviewed publications on a wide range of topics, as well as an inventor on several patents. His current work is related to low data rate access for the Internet of Things, in particular narrowband IoT for LTE with focus on indoor coverage, low cost, long battery life, and large number of devices.

Simone Barbera received his M.Sc.E.E. from the University of Messina, Italy, in January 2006, and his Ph.D. in telecommunications and microelectronics from the University of Rome (Tor Vergata), Italy, in April 2011. He was employed as a hardware engineer at Elital Srl, a company located in L'Aquila, Italy. Later he was employed as a postdoctoral researcher at Aalborg University, working in collaboration with Nokia. His work was mainly related to the standardization of LTE-Advanced, with particular focus on mobility in heterogeneous networks. His main expertise is in hardware design, cellular networks, and global navigation satellite systems. Starting in February 2016, he is working at Telenor A/S, Aalborg, Denmark.

Esa Malkamäki (esa.malkamaki@nokia.com) received M.Sc.(Tech), Lic.Sc.(Tech), and D.Sc.(Tech) degrees, all in electrical and communications engineering, from Helsinki University of Technology in 1989, 1992, and 1998, respectively. He worked for the Communications Laboratory of Helsinki University of Technology participating in the RACE Mobile project starting in 1988. He joined Nokia Research Center in 1992 working first in the RACE ATDMA project. Since 1998 he has been working with standardization research, first with ETSI (EGPRS) and since 1999 with 3GPP (WCDMA, HSPA, LTE, LTE-A), working first on physical layer and recently on higher layer protocols (MAC, RLC, PDCP, RRC) with carrier aggregation, dual connectivity, license assisted access, and latency reduction. He is currently a senior specialist, radio research at Nokia-Bell Labs and participates in 3GPP RAN2 meetings as a standards delegate.

Tero Henttonen (tero.henttonen@nokia.com) received his M.Sc. in applied mathematics from Helsinki University in 2001. He joined Nokia in 2000, first working on standardization research in UMTS/HSDPA/HSUPA and later on LTE in areas of simulation modeling, system simulations, higher layer protocols (MAC, RLC, PDCP, RRC), and mobility. He joined Renesas Mobile Networks at the end of 2010 and worked there as RAN2 standardization delegate for LTE control plane, RRC, dual connectivity, eICIC/feICIC, and small cells aspects. In October 2013 he joined Nokia Networks as a RAN2 delegate for control plane, ASN.1, RRC, dual connectivity, license-assisted access, and carrier aggregation. He is currently a senior specialist, 3GPP standardization at Nokia-Bell Labs and participates in 3GPP RAN2 meetings as a standards delegate.

Benoist Sébire (benoist.sebire@nokia.com) received his Master's degree in electronics and computer engineering with honors from the École Nationale Supérieure des Sciences Appliquées et de Technologie, France. He is currently a chief architect at Nokia-Bell Labs, Tokyo. He has been an active delegate for the past 16 years in 3GPP: first in TSG GERAN for EDGE, and since 2004 in TSG RAN for HSUPA and LTE. His work is related to radio protocol design, and he is the author/co-author of more than 180 patents. He has held numerous positions in 3GPP: as vice-chairman in RAN2 and as specification and work item rapporteur, notably for the LTE Stage 2, since 2006.

Radio Propagation Modeling for 5G Mobile and Wireless Communications

Jonas Medbo, Pekka Kyösti, Katsutoshi Kusume, Leszek Raschkowski, Katsuyuki Haneda, Tommi Jamsa, Vuokko Nurmela, Antti Roivainen, and Juha Meinilä

The authors identify requirements of 5G radio propagation models for relevant propagation scenarios and link types derived from the analysis of recently discussed 5G visions and respective 5G technology trends. They also present a novel map-based propagation model that satisfies the model requirements, and introduce new extensions to existing stochastic models.

Abstract

This article first identifies requirements of 5G radio propagation models for relevant propagation scenarios and link types derived from the analysis of recently discussed 5G visions and respective 5G technology trends. A literature survey reveals that none of the state-of-the-art propagation models such as WINNER/IMT-Advanced, COST 2100, and IEEE 802.11 fully satisfies the model requirements without significant extensions, and therefore there is room for a new framework of propagation models. We then present a novel map-based propagation model that satisfies the model requirements, and also introduce new extensions to existing stochastic models. Several open issues are finally identified that require further studies in 5G propagation modeling.

Introduction

Recently, there have been various international activities to discuss what the next generation system, that is, fifth generation (5G), will be around 2020 and beyond (e.g., [1, 2]). It is generally predicted that areas of mobile services will be significantly expanded by a wide variety of use cases with challenging and diverse requirements in terms of data rate, number of connections, latency, and energy consumption, among other relevant metrics.

A 5G concept, along with relevant technology components, is being developed to address those future requirements (e.g., [1, 3]). These aspects are also translated to 5G propagation modeling requirements. To achieve higher data rates, radio frequencies above 6 GHz have been attracting attention as one of the promising solutions because of their potential to allow wider bandwidths than legacy radio systems operating below 6 GHz. In particular, ultra-dense networks (UDNs) using small cells can take advantage of the propagation properties of the high frequencies, showing higher path loss in the surrounding environment for improving multi-user and multi-cell interference management over space. Massive multiple-input multiple-output (M-MIMO) is another future technology that uses hundreds of antenna elements to efficiently steer signals to dedicated terminals. M-MIMO is a promising technology at both legacy below-6-GHz and higher frequency bands. In contrast to the existing mobile wireless standards, which have mainly targeted human-centric services, a tremendous amount of data traffic is expected to originate from machine-type communication services leading to massive machine communications (MMC) in the 5G system. Direct device-to-device (D2D) communication is seen as an enabler for MMC and also for cellular traffic offloading, coverage extension (e.g., emergency communications), as well as for latency-critical applications (e.g., remote driving, industry automation, tele-protection, and mission-critical controls). Vehicle-to-vehicle (V2V) communication is one specific example of the D2D communications.

As radio channel models are commonly used to evaluate wireless system performance, especially for new technology components, it is essential to have model frameworks that reproduce radio channel responses as close to reality as possible. Given the 5G context, all the mentioned technical aspects set new requirements for modeling. One of the main contributions of this article is to present these key requirements and propagation phenomena that are needed for the evaluation of 5G systems. Furthermore, an overview of currently existing channel models and their shortcomings is given. We then present a new channel model approach and extensions of the existing ones that resulted from the measurement-based channel modeling work within the METIS project.

5G Propagation Model Requirements

As discussed in the previous section, diverse use cases and requirements are foreseen for 5G, which lead to a wide range of relevant propagation scenarios and link types that have to be modeled. The propagation scenarios include environments such as dense urban, urban, indoor office, shopping mall, rural, highway, and stadium, while different link topologies like outdoor-to-outdoor (O2O), outdoor-to-indoor (O2I), and indoor-to-indoor (I2I) are possible. The link types include cellular access, point-to-point such as backhaul, and peer-to-peer links represented by D2D, MMC, and V2V commu-

Jonas Medbo is with Ericsson Research; Pekka Kyösti is with Anite Telecoms and the University of Oulu; Katsutoshi Kusume is with DoCoMo Euro-Labs; Leszek Raschkowski is with Fraunhofer Heinrich Hertz Institute; Katsuyuki Haneda is with Aalto University; Tommi Jamsa is with Tommi Jämsä Consulting and Huawei Technologies; Vuokko Nurmela is with Nokia Bell Labs; Antti Roivainen and Juha Meinilä are with the University of Oulu.

0163-6804/16/$25.00 © 2016 IEEE

nications. The diverse propagation scenarios and link types set the following requirements of the 5G propagation models in addition to the challenges in their implementation in practice.

Applicability to Dual-Mobility Channels

Involvement of device mobility at two link ends as represented in D2D and V2V communications, which we call dual mobility in this article, incurs unique challenges in the propagation modeling (i.e., making the model spatially consistent). This is equivalent to temporal consistency when a device moves over space as time elapses. The propagation model is spatially consistent if two closely located devices in space see similar radio channel profiles in the angular, delay, power, and polarization domains. The consistency therefore ensures that the channels evolve smoothly without discontinuities when devices move or turn around. The lack of spatial consistency potentially leads to significant errors in evaluating radio networks involving device mobility, including wrong handover decisions, unrealistic multihop scenarios, and so on. Spatially consistent modeling is also crucial in MMC and cellular access links such as the UDN, as the density of links is expected to increase and the devices are spatially close to each other.

Applicability to Frequency-Agile Channels

Design of 5G cellular access links requires the propagation model to cover a wide frequency range from 0.5 to 100 GHz. This range is extremely wide compared to the spectrum discussed, for example, in 2G, 3G, and 4G. Although the propagation characteristics, especially diffraction, scattering, and penetration, show significant differences in attenuation at 100 GHz compared to those at 1 GHz, the propagation model should be consistent and applicable across the whole range.

Applicability to Massive-Antenna Channels

5G cellular communication systems aggressively exploit multiple antenna transmission techniques such as spatial multiplexing and spatial division multiple access. Many of these techniques, like M-MIMO and pencil beamforming, will utilize highly resolved spatial properties of the radio channel. Particularly for high carrier frequencies, in the millimeter-wave (mmWave) range, narrow beams are required in order to compensate for the smaller omni-antenna aperture and also link blockage losses in diffraction at building corners and blocking by human bodies, moving objects, and vegetation. Furthermore, if the array antenna is large in respect to the wavelength, radio signals emanating from nearby wireless devices or scatterers cannot be approximated as plane waves, but have to be treated as spherical waves, which can possibly have an impact on beamforming methods. The knowledge of high-frequency radio channels along with the support of M-MIMO is very relevant in cellular UDNs.

Diversity to Accommodate Different Simulation Needs

Finally, on a practical side, the wide range of propagation scenarios and link types sets a challenge of having a single scalable framework of the propagation model applicable to all the possible envisaged scenarios and link types. A model framework for long-range backhaul links may not be used to characterize indoor D2D channels. Furthermore, requirements of the model vary significantly for different link types. A massive sensor network in the form of D2D links may, for example, be based on very simple transceiver units with one antenna each that would not need an angular-dependent propagation model. On the contrary, when looking at cellular access links exploiting M-MIMO, the angular information is crucial, as described earlier. The more requirements imposed on the model, the more complex the implementation and computation. Therefore, it is important to use the right model framework that satisfies the model requirements with minimal complexity. It may be inevitable to have multiple propagation model frameworks with varying levels of requirements addressed, and hence varying complexity, to serve for different propagation scenarios and link types efficiently.

5G Propagation Modeling Approaches

Existing Models Compared to METIS Models

This section reviews the existing models in the literature to see if the 5G propagation model requirements identified in the previous section are addressed.

WINNER-Family Channel Models: The family of geometry-based stochastic channel models (GSCM) includes WINNER [4], IMT-Advanced [5], and Third Generation Partnership Project (3GPP) stochastic channel model (SCM) and D2D model. Although they were originally designed for 2D propagation, further development has led to 3D extensions like WINNER+ [6], QuaDRiGa [7], and 3GPP-3D [8]. They are versatile models for frequencies below 6 GHz supported by a vast amount of channel measurement campaigns. System-level evaluations are supported by the so-called drop concept, which produces non-correlated channel realizations and also correlated large-scale parameters, like shadowing, and angular and delay spreads, for moving user terminals. Model parameters have been missing for frequencies higher than 6 GHz, which is a problem that is partially addressed by METIS 60 GHz measurements (e.g., [9]).

The GSCM framework has major challenges in satisfying the 5G propagation model requirements. For instance, the widely used WINNER-type channel models do not provide correlated channel realizations even if two user terminals are defined close to each other spatially, and hence, the spatial consistency is not supported. This exaggerates the performance of spatial techniques as in reality, the angular separability of the two links is limited because same clusters[1] are visible to both links, resulting in the small-scale channel characteristics of those links being similar. Moreover, the WINNER family models lack realistic amplitude representation of highly resolved sub-paths, resulting in overestimated performance in the case of M-MIMO. This is illustrated in Fig. 1, where the WINNER type of modeling is compared to a measured channel [10]. The singular value distribution of the WINNER modeling method results in a nearly ideal MIMO channel for the large anten-

5G cellular communication systems aggressively exploit multiple antenna transmission techniques such as spatial multiplexing and spatial division multiple access. Many of these techniques, like M-MIMO and pencil beamforming, will utilize highly resolved spatial properties of the radio channel.

[1] Clusters are defined as groups of radio wave scatterers producing multipath echoes to the receiver.

Features	3GPP SCM	WINNER II/ WINNER+	IMT-Advanced	3GPP D2D	3GPP 3D	COST 2100	IEEE 802.11ad	METIS models	
								Stochastic	Map-based
Frequency range (GHz)	1–3	1–6	0.45–6	1–4	1–4	1–6	60–66	0.45–6, 60–70	Up to 100
Bandwidth (MHz)	5	100	100	100	100	200	2000	100 below 6 GHz, 1000 @ 60 GHz	10% of the center freq.
Support M-MIMO	No	Limited	No	No	No	Yes	Yes	Limited	Yes
Support spherical waves	No	No	No	No	No	Yes	No	No[1]	Yes
Support extremely large arrays beyond consistency interval[3]	No	No	No	No	No	Partly	No	No	Yes
Support dual mobility	No	No	No	Limited	No	No	No	Limited[2]	Yes
Support mesh networks	No	No	No	No	No	No	No	No	Yes
Support 3D (elevation)	No	Yes	No	No	Yes	Partly	Yes	Yes	Yes
Support mmWave	No	No	No	No	No	No	Yes	Partly	Yes
Dynamic modeling	No	Very limited	No	No	No	Yes	Limited	No[1]	Yes
Spatial consistency	No	No	No	No	No	Yes	No	Shadow fading only	Yes

[1] Possible, if the location of the physical scattering object is fixed.
[2] Spatially consistent shadowing, azimuth angle of arrival (AOA)/azimuth angle of departure (AOD)/Doppler.
[3] Consistency interval means the maximum distance that, within the large-scale parameters, can be approximated to be constant.

Table 1. Comparison of existing models with METIS models [10].

na array (an even distribution is optimal as the MIMO singular values correspond to the signal-to-noise ratio, SNR, values of the possible MIMO data transmission streams), whereas the measured channel performs much worse. In order to provide a solid basis for the optimization of M-MIMO transmission techniques for 5G, the corresponding channel modeling needs substantial improvement.

COST 2100 Channel Model: The COST 2100 channel model is better suited for spatially consistent modeling of propagation channels. In contrast to the earlier mentioned model family, the COST 2100 model defines clusters on a coordinate system of the environment simultaneously for all user terminals including those in proximity to each other. Each cluster has a visibility region stretching over a spatial area in the environment and determining whether a user terminal "sees" the cluster. Thus, closely located users experience similar propagation environments. Also, spherical waves and smooth time evolution of the channel are supported because of the coordinate-system-based cluster definition. Still, and similar to the WINNER family models, the COST 2100 model is not applicable to dual-mobility channels since it is designed for conditions where one link end, that is, a base station (BS), is fixed. Moreover, the COST 2100 model has only limited support for propagation scenarios and carrier frequencies below 6 GHz.

IEEE802.11ad Channel Model: The IEEE 802.11ad channel model, for very high data rate WLAN, was developed for frequencies around 60 GHz. The model supports spatio-temporal-polarimetric propagation characteristics of non-stationary channels. Line-of-sight, and first- and second-order reflections are modeled based on accurate environment layouts. Intra-cluster properties associated with each reflection are characterized for 60 GHz and for three indoor scenarios only. The model has limited applicability to dual-mobility channel simulations since the cluster properties changes significantly after major motion of WLAN devices. Moreover, cluster coordinates are not utilized, which prevents spherical wave modeling.

Table 1 summarizes the main features of a set of existing models and the two METIS model alternatives that are introduced in the next sections. The comparison reveals that none of the existing channel models fulfills all the listed features and hence satisfies the 5G model requirements.

METIS Model (I): Map-Based Model

As reviewed above, it is a considerable challenge to fulfill all the 5G requirements by extending the existing stochastic GSCM-family models with new features and parameters. Stochastic distributions of the necessary parameters (about 30 in [4], more than 40 in [10]) for all 5G frequency band and environment combinations must be determined such that the resulting model parameters would be consistent across frequency. In order to provide a reliable model parametrization of such a channel model, a large number of extensive channel measurements corresponding to all the modeled environments would be required, which might not be a viable way for-

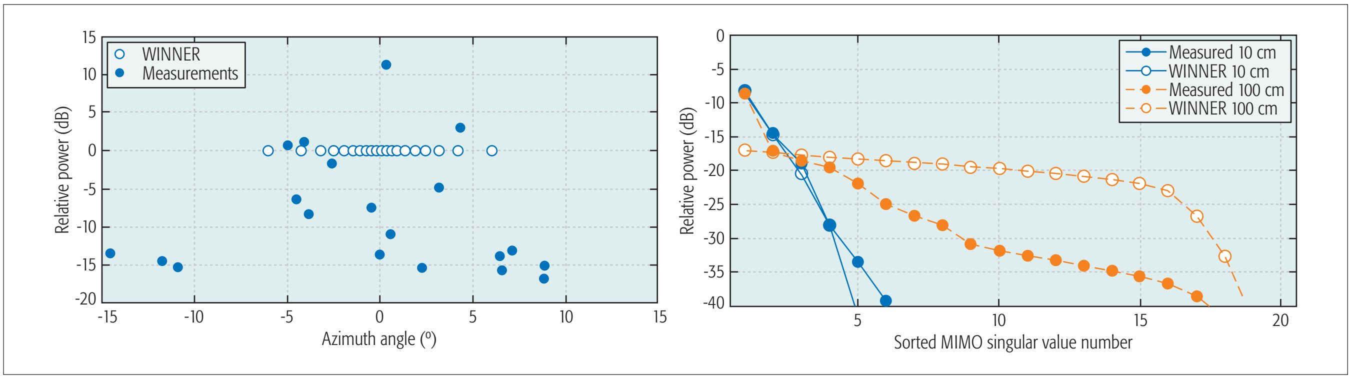


Figure 1. Cluster angle distribution of a real measured urban macro channel and the WINNER model (left) and sorted power distribution of corresponding 40 GHz MIMO channel (20 × 20 elements) singular values for different antenna array lengths of 10 and 100 cm (right).

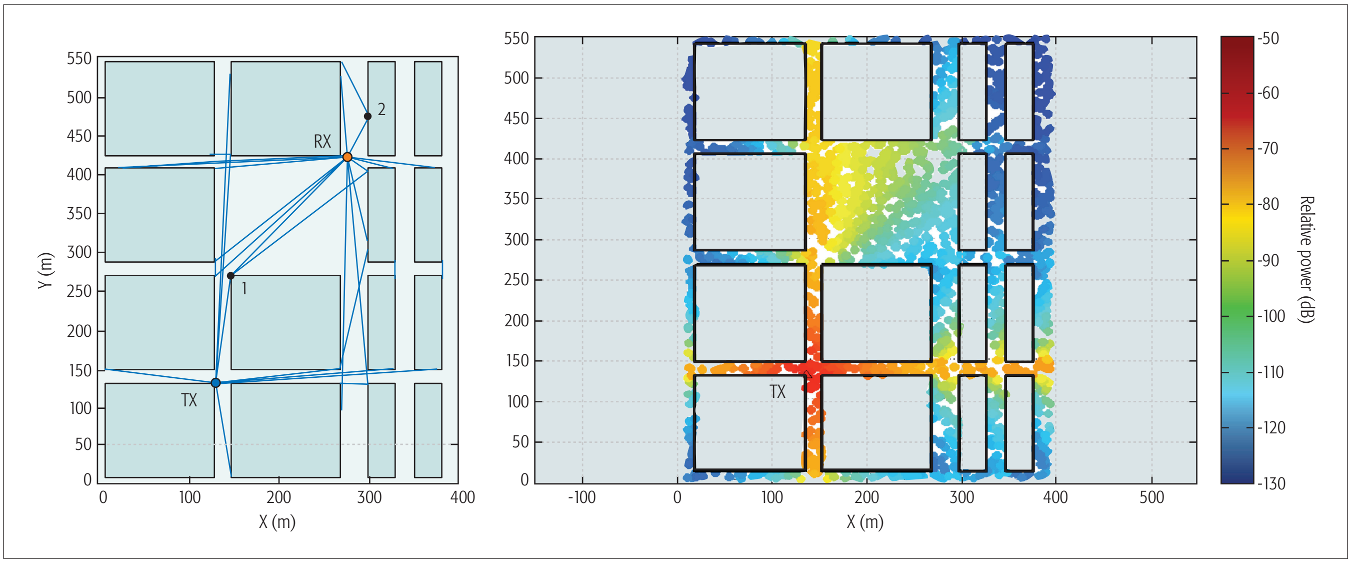


Figure 2. Madrid grid and an illustration of path gains in decibel units. One diffraction point (1) and one specular reflection point (2) are indicated in the left graph.

ward. For this reason METIS provides an alternative modeling approach referred to as the "map-based" model [10].

The map-based model is based on simplified standard ray-tracing techniques with added important features and a coarse geometrical description of the environment. An example of such an environment description is the Madrid grid, depicted in Fig. 2 (also [11]), which has been specified for the METIS test cases. The map-based model inherently addresses all the critical 5G channel modeling challenges as it is based on physical principles using only a limited number of parameters corresponding to the relevant physical properties. In the following a brief overview of the METIS map-based channel model is provided. A detailed description including model parameters is provided in [10]. Notice that the model does not require specific optimization of parameters from measurements for all environment, frequency band, and deployment combinations.

Model Specification: A block diagram of the channel model is illustrated in Fig. 3 with numbered steps of the procedure to generate radio channel realizations. On a higher level the procedure is divided into four main operations: creation of the environment, determination of propagation pathways, determination of propagation channel matrices for path segments, and composition of the radio channel transfer function. In the following we describe the main operations briefly.

Steps 1–4: In the first four steps the 3D propagation environment is specified. The map contains coordinate points of wall corners (e.g., Point 1 in Fig. 2) where for simplicity walls are modeled as rectangular surfaces. Second, a set of random scattering/shadowing objects, representing humans, vehicles, and so on, is drawn on the map with a given scenario-dependent density. Third, rough surfaces (e.g., brick walls) are divided into tiles with certain tile center coordinate points, which act as point sources of diffuse scattering. In Step 4 transceiver locations or trajectories are defined. It is also possible to draw the transceiver locations randomly, which is analogous to drop simulations of GSCMs.

Steps 5–6: The next operation is to determine propagation pathways from the transmitter to the receiver. Coordinates of interaction points for parameter vectors are determined utilizing mathematical tools of analytical geometry. The principles of this part are simple and obvious to

For specular image nodes, blocking occurs also if the path does not intersect the corresponding reflection surface. This procedure may be repeated to achieve any number of diffraction and specular reflection interactions. When repeated, the nodes of previous steps act as TX/RX of the first step.

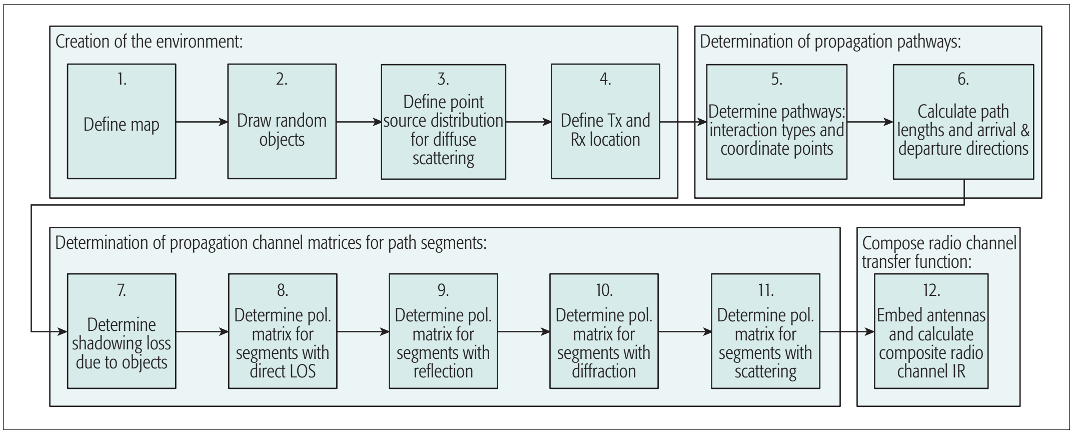

Creation of the environment:
1. Define map
2. Draw random objects
3. Define point source distribution for diffuse scattering
4. Define Tx and Rx location
Determination of propagation pathways:
5. Determine pathways: interaction types and coordinate points
6. Calculate path lengths and arrival & departure directions
Determination of propagation channel matrices for path segments:
7. Determine shadowing loss due to objects
8. Determine pol. matrix for segments with direct LOS
9. Determine pol. matrix for segments with reflection
10. Determine pol. matrix for segments with diffraction
11. Determine pol. matrix for segments with scattering
Compose radio channel transfer function:
12. Embed antennas and calculate composite radio channel IR

Figure 3. A block diagram of the METIS map-based model. Pol: polarization.

the human eye, although writing an algorithmic description of the step is complicated.

Starting from the TX and RX locations (Fig. 2), all possible second nodes visible to the TX/RX node either with a line-of-sight (LOS) path or via a single specular reflection are identified. Possible second nodes are diffraction points like corners, scattering objects, or diffuse scattering point sources. Specular images are also considered as second nodes in this step. Then the coordinates and interaction types of interaction points (diffraction nodes and specular reflection points e.g., Points 1 and 2 in Fig. 2) are determined. Possible pathways are identified by checking whether any wall is blocking the direct or single order reflected paths. For specular image nodes, blocking also occurs if the path does not intersect the corresponding reflection surface. This procedure may be repeated to achieve any number of diffraction and specular reflection interactions. When repeated, the nodes of previous steps act as TX/RX of the first step.

After the pathways are determined, the corresponding path lengths and arrival and departure directions are calculated. The mentioned directions are utilized in the very last step as arguments to radiation patterns of TX and RX antennas.

Steps 7–11: In Step 7 the shadowing due to objects (e.g., humans and vehicles) obstructing or blocking paths is modeled. The blocking effect may be substantial, particularly for higher frequencies in the mmWave range. This effect is accounted for using a simplified blocking model [10]. Each blocking object is approximated by a rectangular screen, as illustrated in Fig. 4. The screen is vertical and, to avoid using multiple screens for each object, perpendicularly oriented with respect to the line connecting the two nodes of the link in the projection from above as shown in Fig. 4. This means that as either node is moving, the screen turns around a vertical line through the center of the screen so that it is always perpendicular to the line connecting TX and RX. Furthermore, each object also scatters the radio waves of nearby paths. The effect of such scatterers is significant when they are located close to either end of the link (TX or RX antennas). It is also significant for scatterers that are in LOS relative to two nodes of a pathway segment, which in turn are in non-LOS relative to each other. For this scattering a simple model of the radar cross-section of a conducting sphere is used [4]. The area A of the screen and the radius of the sphere R are related by $A = \pi R^2$.

For each path segment, propagation matrices are determined for corresponding interactions as indicated in Steps 8–11. The output of these steps is a set of complex 2 × 2 matrices describing gains of polarization components. For example, for the LOS path the matrix is a diagonal matrix with phases and amplitudes based on path length, wavelength, and free space loss. With specular reflection the matrix is determined based on well-known Fresnel reflection coefficients.

The map-based model provides two options for modeling of diffraction. The first option is based on the uniform theory of diffraction (UTD) and provides accurate modeling. A drawback of the UTD approach, however, is that it brings high complexity. For this reason a substantially simpler approach, based on the Berg recursive model [12], is provided as the baseline alternative. The Berg recursive model is semi-empirical and designed for signal strength prediction along streets in an urban environment. It is semi-empirical in the sense that it reflects physical propagation mechanisms without being strictly based on electromagnetics theory. It is based on the assumption that a street corner appears like a source of its own when a propagating radio wave turns around it. The corners of buildings and the antennas represent nodes.

Step 12: The last operation is to compose the radio channel transfer functions by embedding antenna radiation patterns to shadowing losses (from Step 7) and composite propagation matrices. For a single path the complex gain is calculated as a product of the polarimetric antenna radiation pattern vectors, element-wise product of propagation matrices of each path segment of the path, and the total shadowing loss. The result contains all modeled antenna and propagation effects in the given environment for the specified RX and TX antenna locations.

Outdoor to Indoor and Indoor Modeling: For indoor propagation the same ray tracing technique as for outdoors is used with the exception that wall penetration is allowed. There are two complexity levels of determining the indoor penetration loss. For low complexity the loss is mod-

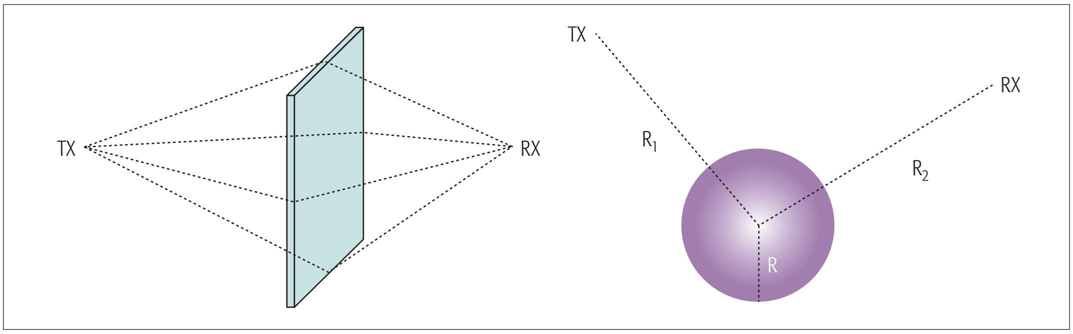


Figure 4. Illustrations of the shadow modeling (left) and scattering modeling (right) of objects of the propagation environment.

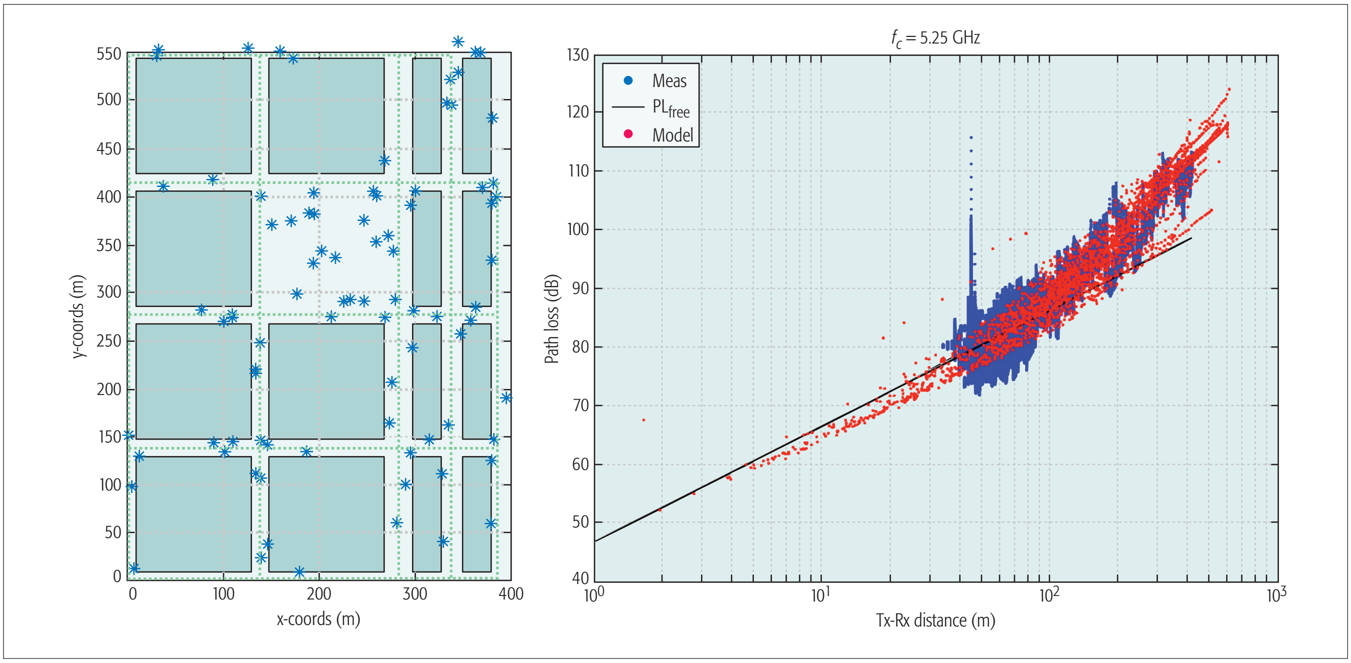


Figure 5. Layout with TX locations denoted by blue stars and RX locations denoted by green dots (left). Comparison of modeled and measured V2V LOS path loss data measured by Oulu University at 5.25 GHz (right).

eled as a constant per unit indoor propagation length (typically in the range 0.3–1 dB/m). For higher complexity a specific loss is assigned to each penetrated indoor wall and/or floor.

For outdoor to indoor propagation a simplified principle is used. The reasoning for the simplification is to keep complexity as low as possible and avoid defining any detailed exterior wall structures such as windows. The model is divided into two cases depending on the level of available detail:

- There is no indoor layout.
- The indoor layout is specified.

In both cases the paths are determined assuming that the building where the user is located does not exist. In other words, the exterior walls are fully transparent in the outdoor-to-indoor direction in the phase of determining propagation paths. When the paths have been identified the building is reintroduced, and the corresponding attenuations for each path due to exterior wall penetration and indoor penetration, as specified above, are determined. To keep the model simple, paths diffracted by, for example, window frames are neglected.

Validation by Measurements: Comparisons with measurement data are crucial to provide validation and reliability of any model. The METIS map-based model has been compared to selected measurement data for this purpose. One example is D2D propagation, which is simulated with the layout of Fig. 5 (left). For this scenario the modeling of scattering and blocking by objects has been successfully validated using two sets of measurement data. The Doppler characteristics, caused by objects along the route of the UE in an urban street, are validated by measurements in [13]. Path loss and shadowing characteristics of the LOS links are shown in Fig. 5 (right) for both the simulation scenario and corresponding measurements in the city of Oulu, which were conducted by University of Oulu and are reported in [10, 14]. The antenna heights are 2.5 and 1.6 m at the different link ends. The frequency is 5.25 GHz. In the model all random objects have the same height of 1.5 m. Thus, no object is fully blocking the direct path. In the measurement higher vehicles were occasionally present, which might have temporarily obstructed the LOS. The spike in measurement results at 40 m is caused by a double-decker bus which blocks the LOS. For this scenario the agreement between measurements and model is evident in spite of the slightly different antenna and obstruction heights in the measurements compared to the model.

The advantage with the map-based model is that it does not need to be validated for all thinkable propagation scenarios. What is important is to validate each model component corresponding to each specific physical propagation mechanism. This validation was only partly performed within the framework of the METIS project. However, it is straightforward to utilize publicly available measurement results to complete the validation of the METIS map-based model.

What is important is to validate each model component corresponding to each specific physical propagation mechanism. This validation was only partly performed within the framework of the METIS project. However, it is straightforward to utilize publicly available measurement results to complete the validation of the METIS map-based model.

METIS Channel Model (II): Stochastic Model Extension

As detailed previously in this article, the stochastic model refers to models based on the GSCM approach, in which scenario-specific parameter distributions are extracted from channel measurements. Model parameter extraction for new scenarios (e.g., moving networks, stadium, UDN, and new frequencies above 6 GHz) is a crucial aspect of fulfilling the requirements of the channel model for 5G simulations. Thus, the METIS stochastic model extension especially focuses on modeling three-dimensional spatial channels in urban microcellular environments, dense urban small cell scenarios, and short-range indoor and outdoor 60 GHz channels (e.g., [9, 10, 15, 16]). The extension includes the following [10] (also see Table 1):

- New frequency agile path loss model for UMi street canyon scenarios covering a frequency from 0.8 to 60.4 GHz
- Model parametrization at 60 GHz in shopping mall [9] and open square scenarios
- Generation of large-scale parameters based on the sum-of-sinusoids method in order to support spatial consistency in the case of moving transmitters and receivers
- Direct sampling of the Laplacian shaped angular spectrum in order to support very large array antennas
- Explicit placing of scattering clusters between TX and RX locations in order to allow for spherical wave modeling to be used

Each of these features was established and supported based on the evidence obtained through extensive channel measurements; the details of the measurements and evidence can be found in [10].

Summary and Future Work

This article introduces a new set of 5G propagation models that are applicable to propagation scenarios and link types derived from recent discussions on 5G visions and respective 5G technology trends. Through the literature survey it is concluded that none of the existing channel models is fully applicable to 5G link design and that consequently new channel models are needed. We present a new map-based model that accounts for all the requirements of 5G propagation model. A brief overview of the new extensions for stochastic models is also provided.

As future work, the model should be validated and reinforced for an even wider range of frequency bands, environments, and network deployment scenarios. Industrial environments for MMC are one of the important scenarios that have been scarcely covered. The literature survey indicated that radio channel measurement results between 6 and 60, and above 70 GHz are far from comprehensive in general. Additionally, most channel sounding has been performed at a single frequency band at different measurement sites. Open questions still remain on a frequency dependent model of diffuse scattering, material absorption, cluster properties, and so on. Finally, it is also intriguing to consider a hybrid approach of the map-based and stochastic models to take advantage of the strength of both models. For example, detailed behaviors of channels (e.g., polarization) are modeled in a physically meaningful manner in the map-based model, but without much comparison with measurements. The stochastic model, on the other hand, is based fully on empirical analysis, while its physical basis is justified only intuitively. One of the possible hybrid approaches of these two models is to add measurement evidence into the physically sound map-based model.

Acknowledgments

The authors would like to thank all our colleagues in METIS, although the opinions expressed herein are those of the authors.

References

[1] METIS, Mobile and Wireless Communications Enablers for the Twenty-twenty Information Society, EU 7th Framework Programme project, http://www.metis2020.com.

[2] NGMN Alliance, "NGMN 5G White Paper," Feb. 2015.

[3] ICT-317669 METIS Project, "Final Report on Architecture," deliverable D6.4, Jan. 2015.

[4] WINNER II D1.1.2, "Channel Models," v. 1.2, 2008.

[5] ITU-R M.2135-1, "Guidelines for Evaluation of Radio Interface Technologies for IMT-Advanced," tech. rep., Dec. 2009.

[6] WINNER+ D5.3, "Final Channel Models," v. 1.0, CELTIC CP5-026 WINNER+ project, http://projects.celtic-initiative.org/winner+/deliverables_winnerplus.html, 2010.

[7] S. Jaeckel *et al*., "QuaDRiGa: A 3-D Multicell Channel Model with Time Evolution for Enabling Virtual Field Trials," *IEEE Trans. Antennas and Propagation*, 2014.

[8] 3GPP TR 36.873, "Study on 3D Channel Model for LTE," v. 12.2.0, June 2015.

[9] A. Karttunen *et al*., "Radio Propagation Measurements and WINNER II Parametrization for a Shopping Mall at 61–65 GHz," *Proc. VTC 2015-Spring*, May 2015.

[10] ICT-317669 METIS Project, "METIS Channel Models," deliverable D1.4 v. 3, June 2015.

[11] ICT-317669 METIS Project, "Simulation Guidelines," deliverable D6.1, Nov. 2013.

[12] J.-E. Berg, "A Recursive Method for Street Microcell Pathloss Calculations," *Proc. IEEE PIMRC '95*, vol. 1, 1995.

[13] ICT-317669 METIS Project, "Initial Channel Models Based on Measurements," deliverable D1.2, Apr. 2014.

[14] A. Roivainen *et al*., "Vehicle-to-Vehicle Radio Channel Characterization in Urban Environment at 2.3 GHz and 5.25 GHz," *Proc. IEEE PIMRC*, Washington, DC, Sept. 2014.

[15] J. Medbo *et al*., "Channel Modeling for the Fifth Generation Mobile Communications," *Proc. EuCAP '15*, April 2015.

[16] A. Roivainen *et al.*, "Elevation Analysis for Urban Microcell Outdoor Measurements at 2.3 GHz," *Proc. 1st Int'l. Conf. 5G for Ubiquitous Connectivity*, 2014.

Additional Reading

[1] K. Haneda *et al.*, "Frequency-Agile Pathloss Models for Urban Street Canyons," *IEEE Trans. Antennas and Propagation*, in press.

Biographies

Jonas Medbo is currently holding a position as senior specialist in applied propagation at Ericsson Research, Sweden. He received his Ph.D. degree in particle physics from Uppsala University, Sweden, in 1997. Since 1997 he has been with Ericsson Research focusing on propagation research. He has contributed to widely used channel models like Hiperlan/2 and 3GPP SCM, and is currently focusing on 5G channel measurements in the range 0.5 to 100 GHz and modeling for 3GPP and ITU.

PEKKA KYÖSTI received his M.Sc. in mathematics from Oulu University, Finland. From 1998 to 2002 he was with Nokia Networks working in the field of transceiver baseband algorithms. From 2002 to 2013 he was with Elektrobit and since that with Anite in Oulu. Since 2002 he has been working on radio channel measurements, estimation, and modeling. He has participated in the channel modeling work in the European METIS 2020 and IST-WINNER projects since 2004.

KATSUTOSHI KUSUME received his M.Sc. and Dr.-Ing. degrees from Munich University of Technology in 2001 and 2010, respectively. In 2002 he joined DoCoMo Euro-Labs and is currently manager of the Wireless Research Group. He led the work package in the METIS project on scenarios/ requirements, channel modeling, and testbed from 2013 to 2015. He received the Best Paper Award at IEEE GLOBECOM '09. His research interests include multiple antennas, iterative processing, and waveform designs.

LESZEK RASCHKOWSKI received his Dipl.-Ing. (M.S.) degree in electrical engineering in 2012 from Technische Universität Berlin, Germany. Currently, he is employed as a research associate at Fraunhofer Heinrich Hertz Institute, Berlin. His research interests include measuring, modeling, and simulating radio propagation channels, as well as performance analysis of wireless communication systems.

KATSUYUKI HANEDA is an assistant professor at Aalto University, Finland. He was the recipient of Best Paper Awards in VTC 2013-Spring and EuCAP 2013. He serves as an Associate Editor of *IEEE Transactions on Antennas and Propagation* and an Editor of *IEEE Transactions on Wireless Communications*. His research focuses on high-frequency radios, wireless for medical and post-disaster scenarios, radio wave propagation modeling, and in-band full-duplex radio technologies.

TOMMI JAMSA graduated from Oulu University in 1995. During his career at Elektrobit (EB) and Anite Telecoms (1993–2015), his responsibilities have been product management, radio channel research, and standardization. He has contributed channel models and test methodologies to several international fora and projects such as COST, WiMAX, 3GPP, ITU-R, WINNER, and METIS. Currently he is a consultant in Tommi Jamsa Consulting, and acts as a senior expert on channel modeling at Huawei Technologies, Sweden.

VUOKKO NURMELA received her M.Sc. in physics from the University of Helsinki, Finland, in 1998. She has been working for Nokia since 1997. Her professional interests include radio propagation measurements and modeling. She has also been working on development and simulations in 2G, 3G, 4G, and 5G systems. She is currently working in Bell Labs Espoo, Finland.

ANTTI ROIVAINEN received his M.Sc. degree in electrical engineering from the University of Oulu in 2007. He is currently working toward his Dr.Sc. (Tech.) degree at the Centre for Wireless Communications (CWC), University of Oulu. He has been involved in several radio channel measurement and modeling activities including modeling of terrestrial and satellite radio channels. His research interests include radio channel measurements and modeling as well as performance analysis of hybrid satellite-terrestrial systems.

JUHA MEINILÄ received his M.S. degree in electrical engineering from the Technical University of Helsinki in 1979. He has worked at the Radio Department of the Finnish PTT and in the Research Center of Finland (VTT). From 1998 to 2015 he worked at Elektrobit, where he focused mainly on development and research. At Elektrobit he participated in the METIS project phase I (2012–2015). In METIS he has participated mainly in channel modeling.

Physical-Layer Authentication for Wireless Security Enhancement: Current Challenges and Future Developments

Xianbin Wang, Peng Hao, and Lajos Hanzo

The authors review the existing physical-layer authentication techniques and identify their current limitations, ranging from low authentication reliability to the difficulties of integrating these techniques with the existing wireless infrastructure and applying them in complex future networks. They then present three promising research areas in addressing these challenges. Specifically, they propose the use of the multi-attributes multi-observation technique for enhancing authentication reliability.

ABSTRACT

While the open nature of radio propagation enables convenient "anywhere" wireless access, it becomes the root of security vulnerabilities in wireless communications. In light of this, physical-layer authentication, which is based on exploitation of the dynamics of physical layer attributes, is emerging as an effective approach to enhancing wireless security. In this article, we first review the existing physical-layer authentication techniques and identify their current limitations, ranging from low authentication reliability to the difficulties of integrating these techniques with the existing wireless infrastructure and applying them in complex future networks. We then present three promising research areas in addressing these challenges. Specifically, we propose the use of the multi-attribute multi-observation technique for enhancing the authentication reliability. In order to apply point-to-point physical-layer authentication techniques into existing wireless networks, we propose a cross-layer authentication approach relying on a composite security key that can seamlessly integrate physical-layer and upper-layer authentication schemes. We also discuss possible ways of invoking physical-layer authentication to reduce both the complexity and latency of the security processes in complex heterogeneous networks with the aid of the proposed physical security context sharing.

INTRODUCTION

Authentication of a wireless device is conventionally handled above the physical layer using key-based cryptography. Although the effectiveness of such techniques has been proven, the security key distribution and management over dynamic wireless networks face a range of emerging problems. The timely sharing of security keys in highly complex networks supporting a large number of mobile and heterogeneous devices is becoming a new challenge. On one hand, the high computational cost of key generation/detection may result in excessive latencies in large-scale networks, which may become intolerable for delay-sensitive communications. On the other hand, the promise that the digital key cannot be computationally broken still remains mathematically unproven [1]. With the rapid growth of processing power, the time spent on cracking a digital security key could be remarkably shortened. Most importantly, attackers using unauthorized security keys cannot easily be detected when the physical-layer attributes are disregarded, because user identification and access rights are only validated through digital keys.

In contrast to the existing upper-layer security schemes, wireless transmitters can also be validated at the physical layer by verifying the dynamic characteristics of the associated physical communication links and devices [2–6], that is, through physical-layer authentication. The reciprocal channel properties and some of the analog front-end (AFE) imperfections of wireless transceivers primarily constitute two categories of physical-layer attributes for device authentication [2]. Compared to digital key-based authentication, the specific physical-layer attributes are directly related to the communicating devices and the corresponding environment, which are extremely difficult to impersonate. Furthermore, both the channel and device imperfection estimation and compensation techniques constitute inherent functions of communications receivers exploited to improve reception performance. As a benefit of this, physical-layer authentication can be accomplished without incurring additional security overhead.

In this article, we first identify the technical challenges of physical-layer authentication in terms of reliability and integration with the existing network infrastructure and protocols. Three promising directions in overcoming these challenges are discussed. Specifically, we propose to enhance the reliability of physical-layer authentication with the aid of a novel multi-attribute multi-observation (MAMO) technique. Furthermore, we explore the inherent link attributes for physical-layer key generation in enabling the concept of the composite security key (CSK). In doing so, physical-layer authentication can be efficiently integrated with existing cryptography-based infrastructures and protocols. Additionally, the authentication procedure of the future fifth generation (5G) heterogeneous networks may be simplified and enhanced by the proposed predicted physical security context sharing (PSCS).

Xianbin Wang and Peng Hao are with Western University; Lajos Hanzo is with the University of Southampton.

0163-6804/16/$25.00 © 2016 IEEE

Challenges in Physical-Layer Authentication

Disregarding the extensive research attention it has drawn, physical-layer authentication is still far from practical deployment due to several challenges. In this section, three challenges in physical-layer authentication are discussed in detail.

Low Reliability of Physical-Layer Authentication

In general, physical-layer authentication techniques can be classified as channel-based and AFE-imperfection-based approaches, as shown in Fig. 1. Channel-based physical-layer authentication exploits the environment-dependent radiometric features of a specific transceiver pair, such as channel state information (CSI) [3] and the received signal strength indicator (RSSI) [7]. These channel characteristics can be used to differentiate signals arriving from an authorized transmitter and those from spoofing transmitters. However, extensive channel monitoring and frequent adaptation of the authentication rules are required when the channel is non-stationary. This may become a challenge in highly dynamic environments (e.g., vehicle-to-vehicle communications) and sleep-mode-aided networks (e.g., IEEE 802.15.4 networks).

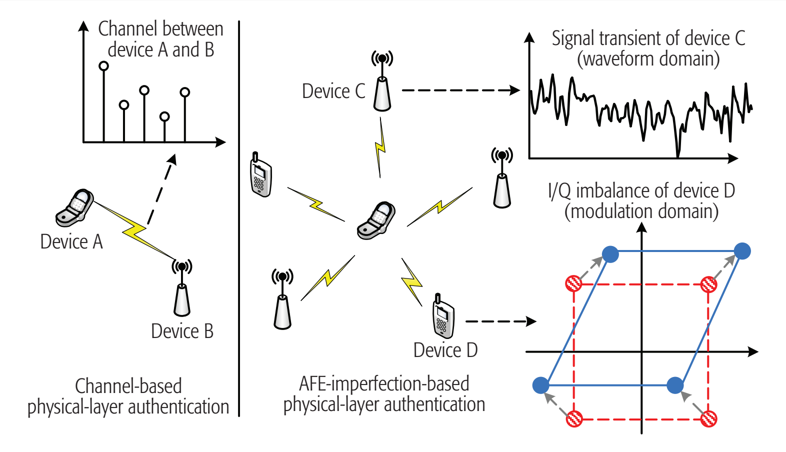
On the other hand, the attributes of the AFE may also be explored for authentication due to its relatively stable nature. These AFE imperfections are inevitable variations introduced to different devices during the fabrication of analog components. Several device-specific characteristics, including the in-phase/quadrature imbalance (IQI) [4], the digital-to-analog converter and the power amplifier characteristics [5], as well as the carrier frequency offset (CFO) [6], have been explored for authentication. In practice, the difference of the selected hardware attributes between different devices is usually small, and its observation is further corrupted by both noise and interference, which reduces the accuracy of estimating these attributes for authentication purposes.

Integration with the Existing Network Infrastructure and Authentication Protocols

Given the significant advantages of physical-layer authentication, it is straightforward to consider the integrated exploitation of physical-layer authentication as a complement to the upper-layer authentication schemes.

One of the most challenging tasks in cross-layer authentication is the integration of physical-layer authentication with the existing infrastructure and protocols. In [2], an overview of cross-layer authentication by using lower-/physical-layer characteristics is provided. Some of the existing cross-layer schemes are implemented through quantization of physical-layer characteristics for upper-layer verification [8]. Although the authentication is realized at an upper layer, the principles of this kind of method and of classic cryptography are rather different. Hence, using it directly in a cryptosystem will impose additional cost, and it is also likely to produce challenges.

Another related challenge is how to extend device-to-device physical-layer authentication to the more general scenarios of end-to-end authentication. In [9], a physical-layer key generation scheme exploiting the channel reciprocity of the directly connected transmitter and receiver is discussed. However, in large-scale wireless networks, authentication and key exchange usually take place between devices that are not directly linked. In contrast, most current physical-layer authentication procedures are limited to device-to-device authentication, since they rely on the characteristics gleaned by analyzing the direct communication links between the transmitter and receiver. As a result, it is critical to develop an authentication process that is not restricted to the physical layer of two directly communicating devices.



Channel between device A and B
Device A
Device B
Channel-based physical-layer authentication
Device C
Device D
AFE-imperfection-based physical-layer authentication
Signal transient of device C (waveform domain)
I/Q imbalance of device D (modulation domain)

Figure 1. Comparison of existing channel-based and AFE-imperfection-based physical-layer authentication techniques.

Authentication in Complex Heterogeneous Networks

It is anticipated that the operational wireless infrastructure will evolve into the 5G by supporting the dramatically increased tele-traffic. Given the significantly increased network complexity, mobile users will have to frequently switch between different base stations or access points, which results in frequent authentication handover. This situation becomes even more challenging in heterogeneous networks (HetNets). The authentication handover is traditionally based on a cryptographic key and on multiple handshakes, as proposed by the Third Generation Partnership Project (3GPP) in [10]. To seamlessly hand over the entire context, the handover has to involve multiple entities including users, access points (APs), base stations (BSs), and servers. Sophisticated backhaul processing and multiple handshakes have to be involved for information or pairwise key exchanges between these entities. In practice, all of these contribute to unwanted latency. This procedure could take up to hundreds of milliseconds, which is far beyond the latency tolerance of 5G services [11].

Future Development of Physical-Layer Security

In this section, we present three possible solutions to address the identified challenges.

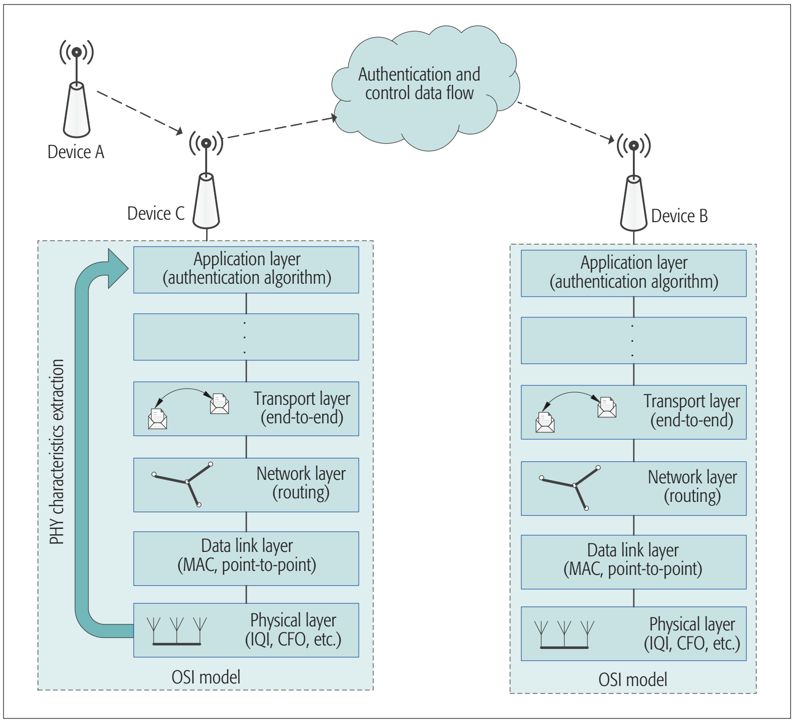

Authentication and control data flow
Device A
Device C
Device B
PHY characteristics extraction
Application layer (authentication algorithm)
Transport layer (end-to-end)
Network layer (routing)
Data link layer (MAC, point-to-point)
Physical layer (IQI, CFO, etc.)
OSI model
Application layer (authentication algorithm)
Transport layer (end-to-end)
Network layer (routing)
Data link layer (MAC, point-to-point)
Physical layer (IQI, CFO, etc.)
OSI model

Figure 2. Cross-layer design for end-to-end authentication.

Reliability Enhancement by Multi-Attribute Multi-Observation Authentication Techniques

As discussed earlier, the performance of physical-layer authentication is often degraded by the instability of the rapidly time-varying channel. Additionally, the performance of AFE-imperfection-based authentication techniques is limited by the low reliability of AFE imperfection estimation.

We propose to enhance the reliability of physical-layer authentication using MAMO techniques by the exploitation of as many of the physical-layer attributes as possible for improving the authentication reliability. Indeed, various channel-based and AFE-imperfection-based physical-layer characteristics may readily be combined for the environment-based characteristics and CSI, as well as some attributes like the RSSI and round-trip time (RRT). As for hardware-imperfection-based characteristics, I/Q amplitude mismatch and phase shift error, the CFO, the clock skew, and so on may be exploited.

The reliability of each physical-layer attribute has to be taken into consideration for multi-attribute-based authentication. The choice of using selected attributes for authentication depends on the specific application scenarios. For instance, the time-invariant AFE imperfections constitute beneficial choices in mobile communications; the channel-based characteristics are expected to work well in stationary indoor scenarios. To elaborate further, we may consider the combination of multiple channel-based and AFE-imperfection-based characteristics for improved authentication performance since it is extremely unlikely for an attacker to occasionally experience the same communication channel and its own nearly identical AFE imperfections as a legitimate transmitter. For example as studied in [12], optimal weights can be set for each of the selected attributes according to their reliability; the authenticity decision can be made either separately or totally based on all the selected characteristics.

The proposed multi-observation technique constitutes another approach to enhancing the characteristic estimation accuracy. Receiver diversity is an effective means of combating wireless fading, which improves the channel capacity by increasing the signal-to-noise ratio (SNR). Given the fact that the estimated characteristics predetermine the attainable authentication reliability, it is plausible that the proposed multi-observation technique improves the authentication reliability. In a cooperative communication system, the source usually relies on multiple relays and optimal relay selection, which facilitates collaborative authentication strategy. For instance, many relays may receive an authentication request from the same source due to the broadcast nature of the wireless medium. Thus, the relays may rely on cooperative observations to jointly authenticate the transmitter for achieving improved authentication reliability.

Seamless Integration with Existing Networks and Protocol Using Composite Security Key

To achieve effective integration of physical-layer authentication and existing network and protocols, two key issues should be considered. First is the proper choice of the physical-layer characteristics that can be extracted for upper-layer security mechanisms. Due to the end-to-end nature of upper-layer authentication, the duration of such a procedure may be significant. Thus, only stable characteristics that are stationary during the authentication process can be exploited. Second, how to process the selected characteristics is another critical concern. The utilization of physical-layer characteristics in widely used symmetric/asymmetric key generation algorithms is an important area for further investigation.

In this subsection, we aim to address the problems in seamlessly integrating the physical-layer and existing upper-layer authentication schemes. We assume that device B needs to authenticate the claimed identity of device A, while devices A and B are in an end-to-end communication scenario, as shown in Fig. 2. Device C, which can be a collaborative access point in practice, is a trusted third party of device B that shares the direct link with A.

The physical layer of our design, which is at the bottom of the protocol stack, plays the critical role of providing characteristics including IQI, CFO, and even antenna-specific characteristics to the upper layers.

The proposed authentication framework is summarized as follows. As a benefit of direct communication with device A, device C becomes capable of evaluating the physical-layer characteristics of A by analyzing its received signals. Therefore, the device-A-specific characteristics can be quantized and hashed at device C for generating specific digital numbers (i.e., PHY-key), which are shared with both devices A and B for further authentication-related processing. Specifically, these PHY-characteristic-related numbers of device A can then be used for generating an asymmetric key for authentication purposes in this article. The PHY-key related to device A can be utilized in the existing key generation algorithm at device A to generate an enhanced asymmetric key pair. The input from the physical layer, actually the PHY-key via device C, can be used as partial input to the private key selection in the existing key generator, thus leading to a physical-layer-dependent composite public and private key pair. On the other hand, the PHY-key can also be directly combined with the original public key from the existing key generator. This will lead to a composite public key capable of preventing unauthorized decryption and cryptanalysis. In this case, additional decryption steps will be required to remove the effect of PHY-key. After generating the composite security keys, the public key can be shared with B with the aid of the existing protocol, while the associated private key is only stored in device A without being shared with any other devices. Basically, device A uses its private key to encrypt plaintext and generate the corresponding ciphertext. Device B attempts to decrypt the ciphertext using the public key, while the authenticity of A is verified only if B is capable of decrypting the readable digest, since only A owns the private key. It is worth noting that PHY-key generation exploiting the hardware-imperfection-related attributes is typically more stable than those gleaned from the wireless channels as argued in [9]. The input from the physical layer expressed in terms of the total number of bits used in the CSK can be adjusted according to the robustness of the physical-layer attributes. In addition, mutual authentication may also be realized through utilization of the shared secret key between A and B with the aid of collaborative devices (e.g., device C for device A).

There are two main benefits of using the proposed PHY-key and composite security key. On one hand, the proposed method could be more efficient. Existing approaches directly using these physical-layer characteristics as an authentication tag will impose additional payload at each layer's data encapsulation, and cost additional bandwidth and power in delivering them to device B. Comparatively, using these characteristics as a securing key can eliminate this overhead. On the other hand, the robustness of the authentication process is enhanced. Similar to the two-factor authentication strategy in which the physical possession factor and virtual password factor are checked together as a double insurance, the PHY-key and CSK are also secured by the intrinsically unforgeable feature of physical-layer characteristics and the computational intractability of asymmetric encryptions.

Authentication Handover Simplification Using Physical Security Context Sharing

In this subsection, we focus on simplifying the authentication procedure in the complex 5G HetNet. The prediction and sharing of physical-layer attributes as security context are the two key aspects of our solution. As illustrated in Fig. 3, we assume a user is moving between cells.

Security Context Prediction: The variation trend of attributes such as direction of arrival (DOA), RSS, RTT, and CSI can be used for physical security context, which can be further predicted based on their previous observations and play an important role in simplifying authentication handover. For example, with the predicted DOA, the authentication-oriented beams of a BS or an AP can accurately point to the antenna array of the intended user, which actively prevents an impersonation attacker from the highly directional communication link between the user and BS/AP. Besides, these attributes can also be used to monitor and track the real-time moving direction and position of the user. The next cell that the user will enter can consequently be predicted. The authentication server thereby is able to prepare the authentication-related information (e.g., the PHY-key information) and send it to the serving AP of the next cell in advance. Once the user enters the new cell, the authentication and association request can be sent back immediately by the serving AP. It is noteworthy that the emerging software defined networking (SDN) can be utilized to efficiently manage the network-wide authentication information as proposed in [13].

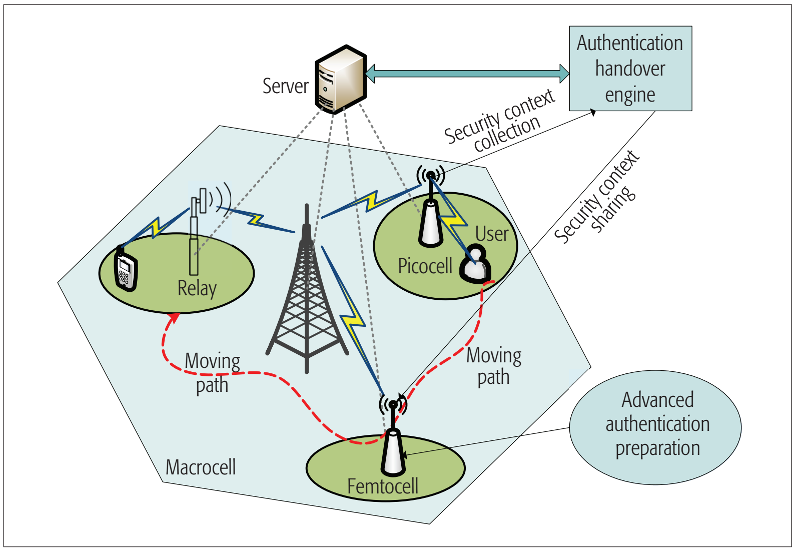

Server
Authentication handover engine
Security context collection
Security context sharing
User
Picocell
Relay
Moving path
Moving path
Advanced authentication preparation
Macrocell
Femtocell

Figure 3. Simplification of authentication handover with physical security context prediction and sharing.

Security Context Sharing: With increased network complexity and operating frequency, more physical characteristics and security context can be observed and shared for authentication purposes. The authentication handover may not happen in a completely new context, implying that much of the already known information of the stable and predictable characteristics can be reused. For example, the PHY-key has high potential to be used as a network-wide unique and unforgeable key because we involve physical-layer factors in key generation. In this case, some repetitive steps such as the frequently repeated pairwise key generation in the solely cryptographic authentication schemes can be reduced.

Case Study and Performance Evaluation

Case Study I: Relay Authentication Using the Multi-Characteristic and Diversity Technique

In the first case study, we consider the authentication of amplify-and-forward (AF) relay as a special case. AF relays, also known as analog repeaters, only work at the physical layer and thus cannot adopt most of the upper-layer authentication schemes. We here apply and evaluate only the physical-layer authentication reliability enhancement using the proposed MAMO technique.

The combination of channel-based RSSI and AFE-imperfection-based IQI may readily be considered as a benefit of their availability in most wireless receivers. The RSSI is representative of the level of the received radio signals and can be directly accessed at the physical layer [7, 14]. Regarding the IQI, we use the same model as in [4], where the AF relay involves one receiving IQI component and one transmission IQI component. For simplicity, we denote the four characteristics of IQI as [α_r, θ_r, α_t, θ_t], where α and θ represent the amplitude and phase shift imbalances, while the subscripts r and t denote reception and transmission. Our objective is to authenticate AF relay nodes by the joint verification of their RSSI and IQI.

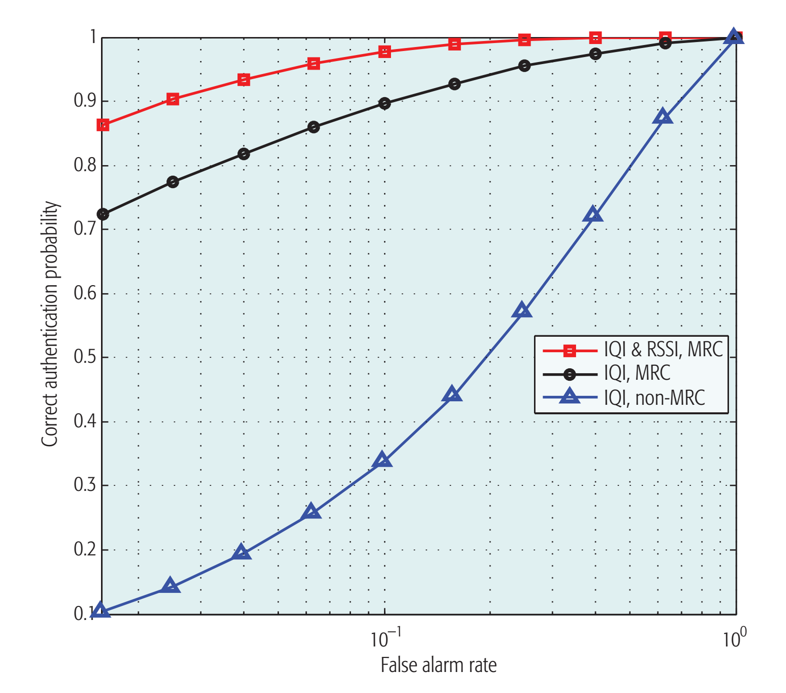


Figure 4. Authentication performance using multiple physical layer attributes.

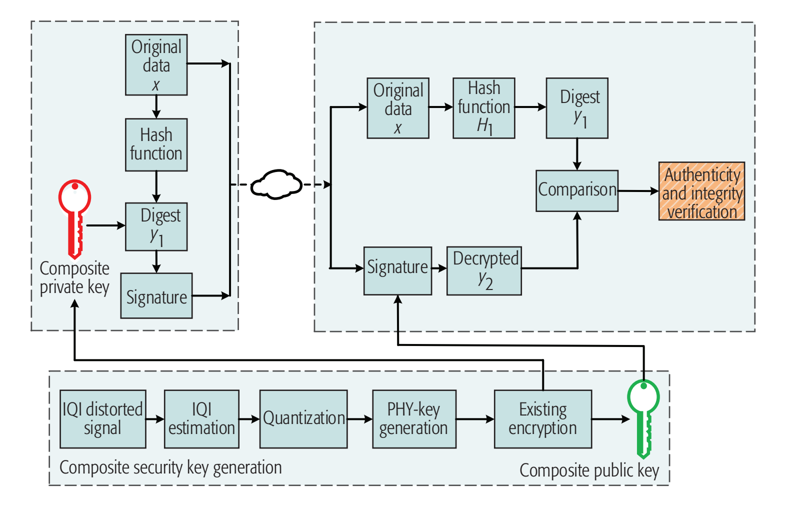


Figure 5. One-way hash digital signature using PHY-key generation.

The proposed authentication procedure is evaluated using MATLAB simulations. We consider four legitimate AF relays and an illegitimate AF relay with α and θ randomly chosen from –0.05~0.05 and –5°~5°, respectively. The RSSI readings of each AF relay are obtained from the experiments using Atheros WiFi devices [14]. We randomly choose one AF relay in each round of simulations, and estimate the corresponding IQI and RSSI of this relay. The generalized likelihood ratio test and hypothesis testing is applied first to determine the relay's authenticity based on the IQI and RSSI separately. Eventually, we combine the two attributes for the final authentication decision. Specifically, we claim to have a legitimate relay only when both the IQI- and RSSI-based tests claim the same legitimate relay nodes. Additionally, three antennas are assumed at the receiver to demonstrate the benefits of multi-observation-based authentication enhancement by using maximal ratio combining (MRC) of multiple received signals. The probability of correct authentication vs. false alarm rate is shown in Fig. 4, where the probability of correct authentication is defined as the percentage of successful authentication trials in the total number of authentication tests. It becomes explicit that the probability of correct authentication is significantly improved by using multiple characteristics (i.e., RSSI and IQI) and MRC-based hypothesis testing. To be specific, the proposed MAMO technique provides on average a 9.28 and 50.48 percent higher correct authentication probability than conventional authentication techniques when only IQI is used in isolation with and without MRC, respectively. This is because the combined characteristics are more reliable and distinguishable than a single characteristic. Additionally, the accuracy of multi-characteristic estimating is further enhanced by combining multiple observations through MRC.

Case Study II: Cross-Layer Authentication Using PHY-Key

In this case study, our proposed cross-layer authentication is evaluated in terms of correct authentication probability and delay reduction.

We first apply the proposed PHY-key into the existing one-way hash digital signature authentication scheme. The block diagram is shown in Fig. 5. For simplicity, we also use IQI to generate the PHY-key in this case study. Without loss of generality, we consider a general transmitter rather than AF relay so that the IQI is modeled as [α_t, θ_t]. As shown in this figure, message-digest 5 (MD5) is used as the hash function to process the quantized IQI and generate the 128-bit hash value; the Rivest, Shamir, and Adleman (RSA) algorithm is used to process the outputs of the compositing procedure in order to generate the public and private keys as presented previously. The MRC relying on multiple antennas is also considered at the receiver to increase the IQI estimate-to-noise ratio.

Additionally, we also simulated the proposed authentication simplification in a handover scenario using the PHY-key. For description simplicity, we assume user U moves to a new cell covered by B from the cell covered by A, while A and B are served by server S. The identity of U has been authenticated by A, that is, A has the knowledge of U's identity as either a legitimate user or an impersonation attacker. We also assume an authorized devices list *AUTH* and an attackers list *ATTK* kept at A, B, and S as $(AUTH, ATTK)_A$, $(AUTH, ATTK)_B$, and $(AUTH, ATTK)_S$, respectively. The lists contain the information of identity, PHY-key, and some predicted directions and positions of different users. Our handover procedure with the prediction and reuse of these lists is presented in Algorithm 1.

1) Start of the authentication handover procedure.
2) A shares the (AUTH, ATTK)A about U to B directly or via S. B updates (AUTH, ATTK)B.
3) U sends B the association request with claimed identity and signature using the above-mentioned one-way hash method.
4) B first checks AUTHB. If U is in AUTHB, B uses the corresponding public-key to decrypt the received signature. If it can decrypt correctly, go to step 7); if it is incorrect, go to step 5). If U is not in AUTHB, go to step 5).
5) B generates PHY-key of U and checks (ATTK)B. If U is in (ATTK)B, go to step 7). If U is not in (AUTH, ATTK)B, B sends S the PHY-key of U, then go to step 6).
6) If S decides to grant U the access, go to 7); otherwise, go to 8).
7) B grants U the access in the response, and go to 9).
8) B rejects U in the response.
9) B shares the updated (AUTH, ATTK)B about U to the next possible cell based on the prediction.
10) End of authentication handover.

Algorithm 1. Authentication handover using PHY-key.

The simulation results on authentication probability vs. SNR and handover delay are shown in Fig. 6. We can see that the probability of correct authentication of our cross-layer authentication increases with the SNR and can be further improved using MRC. It can also be observed that the correct authentication probability is higher than 97 percent even when the SNR is as low as 10 dB, which is representative of a relatively poor wireless communication scenario. For characterizing handover latency, we simulate our handover simplification method and compare it to the traditional handover scheme. The traditional handover scheme of [13] relying on [10] proposed by 3GPP is used in this investigation. In traditional handover, a highly authentication-induced processing delay is imposed by the handover-related request, response, and handshake procedures. In this figure, it can be seen that the delay of both methods increases when network utilization rate (NUR) becomes higher, where NUR is defined as the ratio of actual network traffic to the maximum traffic that the network can handle. We can observe that the handover delay for both methods stays low if NUR is below 60 percent. When the network load becomes high, our method shows its superiority in reducing the delay. Compared to the traditional method, the delay is reduced as we pre-share the (*AUTH*, *s*) of a user by relying

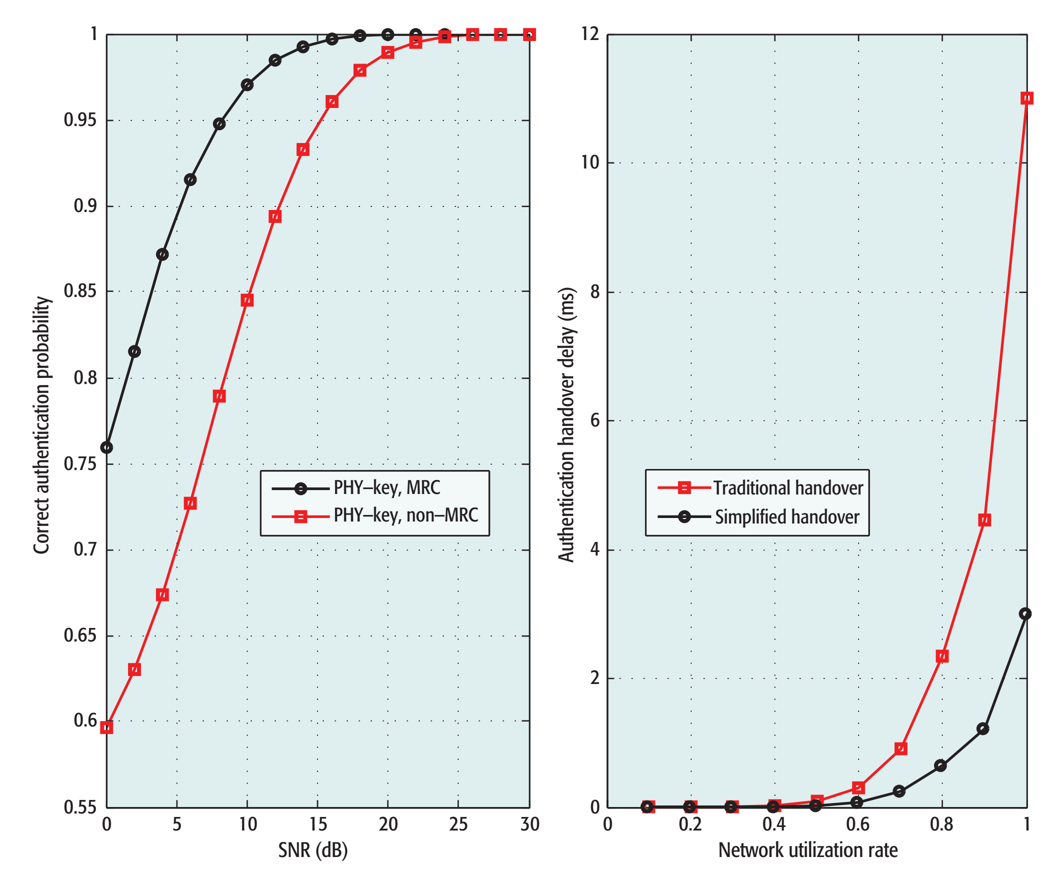


Figure 6. Probability of correct authentication using PHY-key and handover delay.

on the prediction at step 2 and reuse the shared information at step 4.

Conclusions

This article focuses on the current challenges and future development of physical-layer authentication techniques. We identify three main challenges of physical-layer authentication development in terms of the relatively low authentication reliability, seamless integration with existing upper-layer authentication protocols, and the increased authentication complexity problem in 5G. We then propose three solutions to deal with these problems. Specifically, we propose MAMO techniques to enhance the reliability of physical-layer authentication. Also, we propose the cross-layer-aided architecture as well as PHY-key and composite security key generation to achieve seamless integration of physical-layer authentication and cryptography schemes. It is noteworthy that the brute-force search attack, which is the weakness of traditional cryptography, can be alleviated effectively by using the PHY-key. In addition, the upcoming 5G will have fundamental impact on current physical-layer authentication due to increased network complexity. New security approaches including the proposed physical-layer security context prediction and sharing have been studied in simplifying authentication handover in 5G.

The upcoming 5G will bring fundamental impacts to current physical-layer authentication due to increased network complexity. New security approaches including the proposed physical layer security context predication and sharing has been studied in simplifying authentication handover in 5G.

References

[1] A. Mukherjee *et al.*, "Principles of Physical Layer Security in Multiuser Wireless Networks: A Survey," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 3, 2014, pp. 1550–73.

[2] K. Zeng, K. Govindan, and P. Mohapatra, "Non-Cryptographic Authentication and Identification in Wireless Networks," *IEEE Wireless Commun.*, vol. 17, no. 5, Oct. 2010, pp. 56–62.

[3] L. Xiao *et al.*, "Using the Physical Layer for Wireless Authentication in Time-Variant Channels," *IEEE Trans. Wireless Commun.*, vol. 7, no. 7, 2008, pp. 2571–79.

[4] P. Hao, X. Wang, and A. Behnad, "Relay Authentication by Exploiting I/Q Imbalance in Amplify-and-Forward System," *Proc. IEEE GLOBECOM*, Dec. 2014, pp. 613–18.

[5] A. C. Polak, S. Dolatshahi, and D. L. Goeckel, "Identifying Wireless Users via Transmitter Imperfections," *IEEE JSAC*, vol. 29, no. 7, 2011, pp. 1469–79.

[6] W. Hou *et al.*, "Physical Layer Authentication for Mobile Systems with Time-Varying Carrier Frequency Offsets," *IEEE Trans. Commun.*, vol. 62, no. 5, 2014, pp. 1658–67.

[7] Y. Chen *et al.*, "Detecting and Localizing Identity-Based Attacks in Wireless and Sensor Networks," *IEEE Trans. Vehic. Tech.*, vol. 59, no. 5, 2010, pp. 2418–34.

[8] B. Vladimir *et al.*, "Wireless Device Identification with Radiometric Signatures," *Proc. ACM Int'l. Conf. Mobile Computing and Networking*, 2008, pp. 116–27.

[9] K. Zeng, "Physical Layer Key Generation in Wireless Networks: Challenges and Opportunities," *IEEE Commun. Mag.*, vol. 53, no. 6, June 2015, pp. 33–39.

[10] 3GPP TS 33.401 V11.5.0 Tech. Spec. Group Service and System Aspects, "3GPP System Architecture Evolution (SAE); Security Architecture (Release 11)," 2012.

[11] D. He *et al.*, "Secure and Efficient Handover Authentication Based on Bilinear Pairing Functions," *IEEE Trans. Wireless Commun.*, vol. 11, no. 1, 2012, pp. 48–53.

[12] X. Duan and X. Wang, "Authentication Handover and Privacy Protection in 5G HetNet using Software-Defined Networking," *IEEE Commun. Mag.*, vol. 53, no. 4, Apr. 2015, pp. 28–35.

[13] P. Hao, X. Wang, and A. Refaey, "An Enhanced Cross-Layer Authentication Mechanism for Wireless Communications based on PER and RSSI," *Proc. IEEE Canadian Wksp. Info. Theory*, June 2013, pp. 44–48.

[14] P. Hao and X. Wang, "Performance Enhanced Wireless Device Authentication Using Multiple Weighted Device-Specific Characteristics," *Proc. IEEE China Summit & Intl. Conf. Signal and Info. Procesing*, July 2015, pp. 438–42.

Additional Reading

[1] J.G. Andrews *et al.*, "What Will 5G Be?" *IEEE JSAC*, vol. 32, no. 6, 2014, pp. 1065–82.

Biographies

XIANBIN WANG [SM] (xianbin.wang@uwo.ca) is a professor and Canada Research Chair at Western University, London, Ontario, Canada. He received his Ph.D. degree in electrical and computer engineering from the National University of Singapore in 2001. Prior to joining Western, he was with Communications Research Centre Canada (CRC) as a research scientist/senior research scientist between July 2002 and December 2007. From January 2001 to July 2002, he was a system designer at STMicroelectronics, where he was responsible for system design for DSL and Gigabit Ethernet chipsets. His current research interests include 5G networks, communications security, adaptive wireless systems, and locationing technologies. He has over 250 peer-reviewed journal and conference papers, in addition to 24 granted and pending patents and several standard contributions. He is an IEEE Distinguished Lecturer. He has received many awards and recognition, including the Canada Research Chair, CRC President's Excellence Award, Canadian Federal Government Public Service Award, Ontario Early Researcher Award, and three IEEE Best Paper Awards. He currently serves as an Editor/Associate Editor of *IEEE Wireless Communications Letters*, *IEEE Transactions on Vehicular Technology*, and *IEEE Transactions on Broadcasting*. He was also an Editor of *IEEE Transactions on Wireless Communications* between 2007 and 2011. He has been involved in a number of IEEE conferences including GLOBECOM, ICC, VTC, PIMRC, WCNC, and CWIT, in different roles such as Symposium Chair, tutorial instructor, Track Chair, Session Chair, and TPC Chair.

PENG HAO [S] (phao5@uwo.ca) received his B.E.Sc. from Qingdao University, China, in 2008, his M.E.Sc. degree from Shandong University, Jinan, China, in 2011, and his Ph.D. degree from Western University in 2015, respectively. He is currently a postdoctoral fellow with the Department of Electrical and Computer Engineering, Western University. His recent research interests are focused on wireless security, particularly physical-layer authentication using device fingerprint.

LAJOS HANZO [F] (lh@ecs.soton.ac.uk) (FREng, Fellow of IET, Fellow of EURASIP) received his degree in electronics in 1976 and his D.Sc. in 1983. In 2009 he was awarded an honorary doctorate by the Technical University of Budapest and in 2015 by the University of Edinburgh. In 2016 he was admitted to the Hungarian Academy of Science. During his 40-year career in telecommunications he has held various research and academic posts in Hungary, Germany, and the United Kingdom. Since 1986 he has been with the School of Electronics and Computer Science, University of Southampton, United Kingdom, where he holds the Chair in telecommunications. He has successfully supervised about 100 Ph.D. students, co-authored 20 John Wiley/IEEE Press books on mobile radio communications totalling in excess of 10,000 pages, published 1500+ research entries at IEEE Xplore, acted as both TPC and General Chair of IEEE conferences, and presented keynote lectures, and has been awarded a number of distinctions. Currently he is directing a 60-strong academic research team, working on a range of research projects in the field of wireless multimedia communications sponsored by industry, the Engineering and Physical Sciences Research Council UK, the European Research Council's Advanced Fellow Grant, and the Royal Society's Wolfson Research Merit Award. He is an enthusiastic supporter of industrial and academic liaison and offers a range of industrial courses. He is also a Governor of the IEEE Vehicular Technology Society. During 2008–2012 he was the Editor-in-Chief of IEEE Press and a Chaired Professor at Tsinghua University, Beijing. His research is funded by the European Research Council's Senior Research Fellow Grant. For further information on research in progress and associated publications please refer to http://www-mobile.ecs.soton.ac.uk. He has 24,000 citations.

Call for Papers
IEEE Communications Magazine
People-Centric Internet of Things

Background

Internet of Things (IoT) is designed to operate in conjunction with and in service of people. Therefore, people can be viewed as an integral part of the IoT ecosystem. Although considerable work has been done in the recent past regarding IoT, many challenges have remained. In fact, most technologies and solutions for accessing real-world information are either closed, platform-specific, or application-specific. Recent efforts to define IoT reference architectures, such as IoT-A, OpenIoT, SENSEI, or FI-WARE, are important steps in the right direction, but they still lack features that are important for people-centric applications, such as adaptability, intuitiveness, and integration capabilities. So, on one hand, there is need to define an IoT architecture that goes beyond vertical solutions by integrating all required technologies and components into a common, open and multi-application platform. On the other hand, there is need to develop a set of common building blocks, middleware and services that can be used to construct people-oriented applications in an open, dynamic and more effective way into smart environments including but not restricted to smart cities, businesses, education and e-health.

This Feature Topic solicits technical papers describing original, previously unpublished research, not currently under review by another conference or journal, pertaining to People-Centric Internet of Things, including architectural aspects, middleware, and applications. It provides a forum for a broad range of unsolicited high quality scientific research papers that meet the criteria of originality, presentation quality and topic relevance. Submissions should clearly identify how they relate to topics under consideration in this special issue. Contributions describing an overall working system and reporting real world deployment experiences are particularly of interest.

This Feature Topic will focus on several topics such as:

- People-IoT Interactions
- Social Network Applications to Mobile Computing
- Context-Aware Applications and Services
- Human in the Loop
- Big Data Analysis in People-centric IoT
- Cloud-based People-centric IoT Applications and Environments
- Security and Privacy
- Prototypes, Field Experiments, Testbeds

Submissions

Articles should be tutorial in nature and written in a style comprehensible to readers outside the specialty of the article. Authors must follow the IEEE Communications Magazine's guidelines for preparation of the manuscript. Complete guidelines for prospective authors can be found athttp://www.comsoc.org/commag/paper-submission-guidelines. It is very important to note that the IEEE Communications Magazine strongly limits mathematical content, and the combined number of figures and tables to six. Manuscript length (introduction through conclusions, excluding figures, tables and captions) should not exceed 4500 words. Manuscripts should be submitted through Manuscript Central at http://mc.manuscriptcentral.com/commag-ieee/ by the manuscript submission deadline. Please select "February 2017/People-Centric IoT" in the drop down menu. For further details, please refer to 'Information for Authors' on theIEEE Communications Magazine web site at http://www.comsoc.org/pubs/commag/sub_guidelines.html.

Important Dates

- Manuscript Submission: June 30, 2016
- Decision Notification: September 15, 2016
- Final Manuscripts Due: November 15, 2016
- Publication Date: February 2017

Guest Editors

Jorge Sá Silva
University of Coimbra, Portugal
sasilva@dei.uc.pt

Pei Zhang
Carnegie Mellon University, USA
peizhang@cmu.edu

Trevor Pering
Google, USA
peringknife@google.com

Fernando Boavida
University of Coimbra, Portugal
boavida@dei.uc.pt

Takahiro Hara
Osaka University, Japan
hara@ist.osaka-u.ac.jp

Nicolas C. Liebau
SAP AG, Germany
nicolas.liebau@gmail.com

ADVERTISERS' INDEX

ADVERTISING SALES OFFICES

Closing date for space reservation: 15th of the month prior to date of issue

NATIONAL SALES OFFICE
James A. Vick
Sr. Director Advertising Business, IEEE Media
EMAIL: jv.ieeemedia@ieee.org

Marion Delaney
Sales Director, IEEE Media
EMAIL: md.ieeemedia@ieee.org

Mark David
Sr. Manager Advertising & Business Development
EMAIL: m.david@ieee.org

Mindy Belfer
Advertising Sales Coordinator
EMAIL: m.belfer@ieee.org

NORTHERN CALIFORNIA
George Roman
TEL: (702) 515-7247
FAX: (702) 515-7248
EMAIL: George@George.RomanMedia.com

SOUTHERN CALIFORNIA
Marshall Rubin
TEL: (818) 888 2407
FAX: (818) 888-4907
EMAIL: mr.ieeemedia@ieee.org

MID-ATLANTIC
Dawn Becker
TEL: (732) 772-0160
FAX: (732) 772-0164
EMAIL: db.ieeemedia@ieee.org

NORTHEAST
Merrie Lynch
TEL: (617) 357-8190
FAX: (617) 357-8194
EMAIL: Merrie.Lynch@celassociates2.com

Jody Estabrook
TEL: (77) 283-4528
FAX:: (774) 283-4527
EMAIL: je.ieeemedia@ieee.org

SOUTHEAST
Scott Rickles
TEL: (770) 664-4567
FAX: (770) 740-1399
EMAIL: srickles@aol.com

MIDWEST/CENTRAL CANADA
Dave Jones
TEL: (708) 442-5633
FAX: (708) 442-7620
EMAIL: dj.ieeemedia@ieee.org

MIDWEST/ONTARIO, CANADA
Will Hamilton
TEL: (269) 381-2156
FAX: (269) 381-2556
EMAIL: wh.ieeemedia@ieee.org

TEXAS
Ben Skidmore
TEL: (972) 587-9064
FAX: (972) 692-8138
EMAIL: ben@partnerspr.com

EUROPE
Christian Hoelscher
TEL: +49 (0) 89 95002778
FAX: +49 (0) 89 95002779
EMAIL: Christian.Hoelscher@husonmedia.com

COMPANY .... PAGE

CURRENTLY SCHEDULED TOPICS

TOPIC	ISSUE DATE	MANUSCRIPT DUE DATE
INTERNET OF THINGS (IOT)	DECEMBER 2016	JUNE 15, 2016
PEOPLE-CENTRIC INTERNET-OF-THINGS	FEBRUARY 2017	JUNE 30, 2016
SUSTAINABLE INCENTIVE MECHANISMS FOR MOBILE CROWDSENSING	MARCH 2017	JULY 15, 2016
FOG COMPUTING AND NETWORKING	APRIL 2017	SEPTEMBER 1, 2016
NETWORK SLICING IN 5G SYSTEMS	MAY 2017	SEPTEMBER 15, 2016

www.comsoc.org/commag/call-for-papers

JUNE 2016

IEEE ComSoc live WEBINAR

Now available on-demand!

InterDigital's Creating the Living Network ™ Webinar Series

Rumble in the IoT Jungle: Emerging Standards Take On Proprietary Solutions

We know the vision – as the IoT industry grows our "things" will become more intelligent and make our lives more efficient by connecting everything from health monitoring devices to traffic lights. In addition to end user benefits, the opportunity to create new revenue streams through monetized data and new applications makes the IoT space attractive for businesses and services. But what connectivity will underpin the IoT?

The mobile industry is characterized by global standards, and various solutions are being pushed by organizations like 3GPP, ETSI and most emphatically oneM2M. But proprietary solutions aren't waiting, and are driving into the market with connectivity solutions that are ready and available now, and could emerge to be a de facto standard if adoption is rapid enough. Who'll win this battle?

The fourth webinar in InterDigital's Creating the Living Network™ webinar series will bring together two IoT experts to debate proprietary solutions vs. global standards in IoT.

Sponsor content provided by: INTERDIGITAL.

free Tutorials Now™

Emerging Concepts and Technologies Towards 5G Wireless Networks

Despite the recent advances in wireless technologies, the wireless community faces the challenge of enabling a further traffic increase of up to 1,000 times in the next 10 years or so, while no customer is willing to pay more for the wireless pipe itself: the so called "traffic-revenue decoupling." Moreover, many experts warn that the low-hanging fruits in wireless research (in information theory, communications theory, and signal processing) have already been collected. Topics in this tutorial will cover:

◆ Fundamental dynamics of cellular communications ◆ 3GPP operation ◆ Key technologies in LTE and LTE-Advanced ◆ Emerging challenges and opportunities in beyond-2020 wireless networks ◆ Bottleneck problems in beyond-2020 wireless networks

In addition, the potential research directions towards coping with the bottleneck problems, especially in the context of radio access network (RAN) will be discussed and the underlying mathematical tools will also be highlighted. Areas to be addressed will include:

◆ Interference-robust PHY ◆ Non-coherent communications ◆ Steerable beam forming ◆ Uplink distributed multi-channel multiple-access ◆ Inter-cell load coordination (ICLC) for non-uniform traffic ◆ Interdisciplinary approaches + RRM ◆ Robust algorithms and protocols ◆ Layer 8 – User-in-the-loop

IEEE ComSoc content sponsored by: ROHDE&SCHWARZ

IEEE ComSoc™
IEEE Communications Society

Limited Time Only at >> www.comsoc.org

For this and other sponsor opportunities
contact Mark David // 732-465-6473 // m.david@ieee.org

INNOVATE FASTER

WITH FIELD-DEPLOYED 5G PROOF-OF-CONCEPT SYSTEMS

In the race to design next-generation wireless technologies, research teams must rely on platforms and tools that accelerate their productivity. Using the NI software defined radio platform and LabVIEW Communications, leading researchers are innovating faster and building 5G proof-of-concept systems to demonstrate new technologies first.

Accelerate your innovation at ni.com/5g

LabVIEW Communications System Design Software, USRP-2943R SDR Hardware

NATIONAL INSTRUMENTS™