October 2017, Vol. 55, No. 10

IEEE Communications Magazine

- Military Communications
- Design and Implementation
- Integrated Circuits for Communications
- Behavior Recognition Based on Wi-Fi CSI
- IoT and Information Processing in Smart Energy

IEEE ComSoc IEEE Communications Society

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

2950
575
5700
4850
350
75
940
150
1660
1000
40
11
12
1
2
3
10
9
8
7
6
4

IEEE COMMUNICATIONS MAGAZINE

OCTOBER 2017, vol. 55, no. 10
www.comsoc.org/commag

Military Communications

Guest Editors: Kevin S. Chan and Frank T. Johnsen

IoT and Information Processing in Smart Energy Applications

Guest Editors: Wei-Yu Chiu, Hongjian Sun, John Thompson, Kiyoshi Nakayama, and Shunqing Zhang

Behavior Recognition Based on Wi-Fi CSI: Part 1

Guest Editors: Bin Guo, Yingying (Jennifer) Chen, Nic Lane, Yunxin Liu, and Zhiwen Yu

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: $71: print, digital, and electronic. $33: digital and electronic. $1001: non-member print.

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*, IEEE, 445 Hoes Lane, Piscataway, NJ 08855-1331.

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 submitted 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. All submissions will be peer reviewed. For further information contact Tarek El-Bawab, Associate Editor-in-Chief (telbawab@ieee.org).

Design and Implementation Series: Mobile Bandwidth Improvement Techniques

Series Editors: Vijay K. Gurbani, Salvatore Loreto, and Ravi Subramanyan

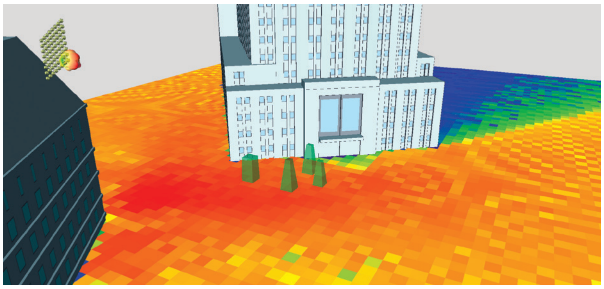
Integrated Circuits for Communications

Series Editors: Charles Chien and Zhiwei Xu

Accepted from Open Call

SCHEDULED TOPICS

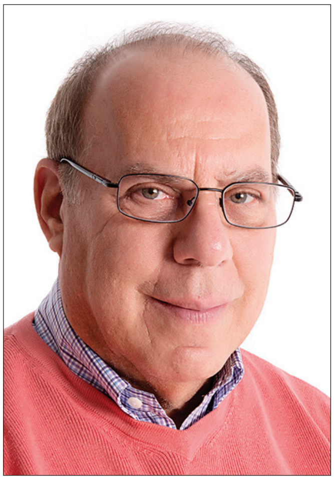
Topic	Publication Date	Submission Due Date
Future 5G Millimeter Wave Systems and Terminals: Propagation Channel, Communication Techniques, Devices and Measurements	June 2018	22 Oct 2017
Orchestration of Ultra-Dense 5G Networks	June 2018	22 Oct 2017
Information-Centric Networking Security	July 2018	1 Nov 2017
Multi-Access Mobile Edge Computing for Heterogeneous IoT	July 2018	1 Nov 2017
Education & Training: Humanitarian Engineering and Community Engagement in Education	May 2018	1 Dec 2017
Enabling Communication and Networking Technologies for Edge Computing	August 2018	1 Dec 2017
Exploring Caching, Communications, Computing and Security for the Emerging Smart Internet of Things	August 2018	1 Dec 2017



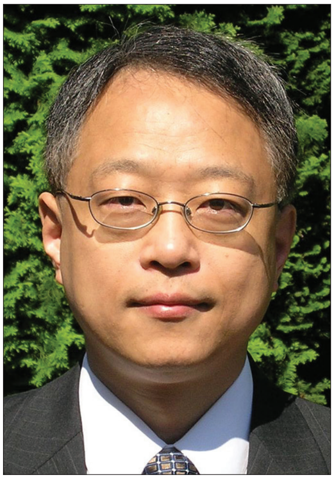
LEVERAGING ADVANCED TECHNOLOGIES

Tao Zhang, the Chief Information Officer (CIO) for the IEEE Communications Society (ComSoc), oversees the planning, development, acquisition, and use of the Society's information systems and applications to ensure effective and efficient support for the Society's services, products, and operations. ComSoc offers a wide range of services and products, including conferences, publications, standardization, and training and education. ComSoc also operates many technical communities, working groups, cross-society initiatives, forums, and meetings. The information systems and applications must also support additional ComSoc operations such as marketing and recruiting. To do all this ComSoc's Information Communications Technology (ICT) must leverage advanced technologies to maximize its effectiveness and efficiency.s

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/CTO for Smart Connected Vehicles. Since then, he has also been driving strategies, technology, and eco-systems for IoT security 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 about 30 years, Tao has been directing research and product development to not just create innovations but also transform them into practical solutions. He is a co-founder and a board director of the OpenFog Consortium, the CIO and a Board Governor of the IEEE Communications Society (2016 and 2017), and a co-founder and a founding Board Director of the Connected Vehicle Trade Association (CVTA). He holds over 50 US patents and has co-authored two books: *Vehicle Safety Communications: Protocols, Security, and Privacy* (2012) and *IP-Based Next Generation Wireless Networks*.



Harvey Freeman



Tao Zhang

First, ComSoc has long desired a way to gain holistic, deeper, and more accurate insight into its members and how its members and others use ComSoc products and services. Who are using ComSoc services and products? How do members and non-members use ComSoc services and products differently? How satisfied are the users? What values and features attract users to ComSoc products and services? What events and factors influence the usage of ComSoc services and products and how? While such information is essential to evaluating existing products and services and developing new ones, obtaining the information had been difficult. The main reason was that different ComSoc services and products (e.g., membership management, conferences, and training and education services) had been using separate systems and tools to collect data, and they often collect inconsistent data elements and store them in different data formats.

ComSoc has developed a ComSoc Member Relational Management System (MRMS) to consolidate data about membership, publications, conferences, and training and education services into a single system. This MRMS provides a holistic and more accurate view of ComSoc members, potential members, and general public users of ComSoc services and products, along with data on how they use ComSoc products and services. It can also make such information available to ComSoc volunteer leaders and staff in a more timely fashion. It further provides easier-to-use tools to perform deeper and more accurate analysis based on the raw data to assist decision making by ComSoc volunteer leaders.

Second, to increase ComSoc's value to its members, ComSoc plans to develop and offer members-only content, products, and services. To enable members-only offerings, we have developed a members-only section (or a members wall) on the ComSoc website (comsoc.org), where members can enjoy contents, products, and services available only to them. We have integrated the ComSoc website with the IEEE Single Sign On system to allow users to log into and navigate both IEEE and ComSoc websites with a single account, and also use this single account to purchase ComSoc products and services. We have further integrated the ComSoc Training E-Commerce system with the IEEE Enterprise Shopping Cart to streamline the process for users to access and purchase ComSoc products and services.

Third, ComSoc conferences have been providing tremendous value to both the academic and industry communities. Essential to the success of these conferences are the tools for the conference organizers to organize and monitor the conferences and to ensure timely publication of the conference papers. We updated the Web content workflow used for running ComSoc conferences to allow conference organizers and ComSoc staff to update conference websites faster and more easily. A long-standing issue the conference organizers have been facing was the lack of ability for them to report "no-show papers" (i.e., papers accepted but not presented by anyone at a conference) at or immediately after a conference. Since the IEEE does not publish no-show papers, the lack of ability to timely report these papers often caused long delays in the publication of conference papers. We have upgraded the Conference Session Chair Report App to enable session monitors to report no-show papers in real time, which allows the presented papers to be published in IEEE Xplore much more quickly than before.

Fourth, to further unify and streamline IT support for the many different products and services ComSoc offers, we have been redesigning the ComSoc website. We are integrating the many separate websites used to support different ComSoc publications (e.g., WCET, Green ICT), ComSoc chapters, and ComSoc committees into the single new website. The new website will offer improved design and functionality, allow easier and faster updates, and further consolidate user and usage data that have previously been scattered among different websites. We have further improved ComSoc website to meet higher Web Accessibility Guidelines (WCAG 2.0 Level A and AA) so potential users can find ComSoc-offered content, services, and products more easily.

Fifth, as users are rapidly shifting to digital editions of ComSoc publications from paper editions, ComSoc has been building the digital editions for all ComSoc publications, with the digitized ComSoc magazines to become available on the new ComSoc website first and soon.

In conclusion, we must ensure that ComSoc continues to leverage advanced technologies to maximize its effectiveness and efficiency. The significant accomplishments by the ICT team over these past two years are just a start and not an end.

Newly Approved Amendments to the IEEE ComSoc Bylaws

On May 24, 2017, the ComSoc Board of Governors approved a revision of the ComSoc Bylaws. The Vice President, Technical Activities IEEE approved the revision on August 23, 2017.

The main changes in this revision can be summarized as follows:

- A change in the organization of the Board of Governors. Two new positions were added: Chief Marketing Officer (CMO) and Director Industry Communities; these are non-voting positions.
- Changed name of Technical Services Board to Technical Committees Board.
- Changed ICEC Standing Committee to Industry Communities Board.
- Elevated Young Professionals Ad-Hoc Committee to a Standing Committee.

The revised Bylaws are now in force and can be found here: http://www.comsoc.org/about/documents/bylaws.

UPDATED ON THE COMMUNICATIONS SOCIETY'S WEB SITE

www.comsoc.org/conferences

2017

OCTOBER

I3C 2017 — IoT Int'l. Innovation Conference, 5–7 Oct.
Saodoa. Morocco
http://i3c2017.emena.org/index.html

IEEE PIMRC 2017 — IEEE Int'l. Symposium on Personal, Indoor & Mobile Radio Communications, 8–13 Oct.
Montreal, Canada
http://pimrc2017.ieee-pimrc.org/2015/08/21/sample-news-post/

IEEE CNS 2017 — IEEE Conference on Communications and Network Security, 9–11 Oct.
Las Vegas, NV
http://cns2017.ieee-cns.org/

HONET-ICT 2017 — Int'l. Conference on Smart Cities: Improving Quality of Life Using ICT & IoT, 9–11 Oct.
Irbid, Jordan
http://honet-ict.org/

WCSP 2017 — Int'l. Conference on Wireless Communications and Signal Processing, 11–13 Oct.
Nanjing, China
http://www.ic-wcsp.org/

CyberC 2017 — Int'l. Conference on Cyber-Enabled Distributed Computing and Knowledge, 12–14 Oct.
Nanjing, China

IEEE HEALTHCOM 2017 — IEEE Int'l. Conference on e-Health Networking, Application & Services, 12–15 Oct.
Dalian, China
http://healthcom2017.ieee-healthcom.org/

IEEE SmartGridComm 2017 — IEEE Int'l. Conference on Smart Grid Communications, 16–19 Oct.
Dresden, Germany
http://sgc2017.ieee-smartgridcomm.org/

CSNet 2017 — Cyber Security in Networking Conference, 18–20 Oct.
Rio de Janeiro, Brazil
http://csnet2017.dnac.org/

ICTC 2017 — Int'l. Conference on Information and Communication Technology Convergence, 18–20 Oct.
Jeju Island, Korea
http://ictc2017.org/

ATC 2017 — Int'l. Conference on Advanced Technologies for Communications, 18–20 Oct.
Quynhon, Vietnam
http://atc-conf.org/

INTEC 2017 — Int'l. Conference on Internet of Things, Embedded Systems and Communications, 20–22 Oct.
Gafsa, Tunisia
http://www.iintec.org/

IEEE/CIC ICCC 2017 — IEEE/CIC Int'l. Conference on Communications in China, 22–24 Oct.
Qingdao, China
http://iccc2017.ieee-iccc.org/

MILCOM 2017 — Military Communications Conference, 23–25 Oct.
Baltimore, MD
http://events.afcea.org/milcom17/public/enter.aspx

Fog World Congress 2017, 30 Oct.–1 Nov.
Santa Clara, CA
http://www.fogworldcongress.com/

NOVEMBER

WINCOM 2017 — Int'l. Conference on Wireless Networks and Mobile Communications, 1–4 Nov.
Rabat, Morocco
http://www.wincom-conf.org/?p=welcome

IEEE NFV-SDN 2017 — IEEE Conference on Network Function Virtualization and Software Defined Networks, 6–8 Nov.
Berlin, Germany
http://nfvsdn2017.ieee-nfvsdn.org/

FRUCT21 2017 — Conference of Open Innovations Association (FRUCT) 2017, 6–10 Nov.
Helsinki, Finland
http://fruct.org/conference21

IEEE LATINCOM 2017 — 9th Latin-American Conference on Communications, 8–10 Nov.
Guatemala City, Guatemala
http://latincom2017.ieee-comsoc-latincom.org/

IEEE COMCAS 2017 — Int'l. Conference on Microwaves, Communications, Antennas and Electronic Systems, 13–15 Nov.
Tel Aviv, Israel
http://www.comcas.org/

IEEE ICSOS 2017 — IEEE Int'l. Conference on Space Optical Systems and Applications, 14–16 Nov.
Naha, Japan
http://icsos2017.nict.go.jp/

ISWSN 2017 — Int'l. Symposium on Wireless Systems and Networks, 19–22 Nov.
Lahore, Pakistan
http://sites.uol.edu.pk/iswsn17/

NOF 2017 — Int'l. Conference on the Network of the Future, 22–24 Nov.
London, United Kingdom
http://www.network-of-the-future.org/

CNSM 2017 — Int'l. Conference on Network and Service Management, 26–30 Nov.
Tokyo, Japan
http://www.cnsm-conf.org/2017/

IEEE VNC 2017 — IEEE Vehicular Networking Conference, 27–29 Nov.
Torino, Italy
http://www.ieee-vnc.org/

ITU-K 2017 — ITU Kaleidoscope: Challenges for a Data-Driven Society, 27–29 Nov.
Nanjing, China
http://www.itu.int/en/ITU-T/academia/kaleidoscope/2017/Pages/default.aspx

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

GCN GLOBAL COMMUNICATIONS NEWSLETTER

October 2017
ISSN 2374-1082

DISTINGUISHED LECTURER TOUR

Distinguished Lecturer Tour of Abbas Jamalipour in Malaysia, 2017

By Aduwati Sali, Fazirulhisyam Hashim, Chee Yen Leow (Bruce), Nur Idora Abdul Razak, and Hafizal Mohamad, IEEE Malaysia ComSoc/VTS Joint Chapter

In April 2017, the IEEE Malaysia Communications and Vehicular Technology Society (ComSoc/VTS) Joint Chapter hosted Prof. Abbas Jamalipour from the University of Sydney, Australia for his Distinguished Lecturer Tour (DLT) at three different venues. His first lecture was given at the Universiti Teknologi Mara (UiTM) Shah Alam (April 11, 2017); the second lecture was at the Malaysian Communications and Multimedia Commission (MCMC) Cyberjaya (April 12, 2017); the third lecture was at the Universiti Teknologi Malaysia (UTM) Skudai, Johor Bahru (April 13, 2017).

Prof. Jamalipour arrived at Kuala Lumpur International Airport 2 on April 9, 2017. He was welcomed by Fazirul, the Past Chair of ComSoc/VTS and brought to the Everly Hotel Putrajaya. On April 10, 2017, Prof. Jamalipour had a lunch meeting and spent time with ComSoc/VTS executive committee members led by Aduwati Sali. Various topics were discussed during this meeting, notably on strategies to strengthen the chapter, potential collaboration, how to attract more members to join IEEE and ComSoc/VTS, etc.

On the morning of April 11, 2017, Prof. Jamalipour delivered his first lecture at the Al-Haytham Seminar Room, Engineering Complex, UiTM, Shah Alam. The title of his lecture was "Wireless Communications Techniques for Emergency and Disaster Situations." There were about 25 attendees, mainly postgraduate students and researchers from various research institutes and industry in Malaysia. In this lecture, Prof. Jamalipour discussed the recent advancements and developments in the field of wireless communications for emergency and disaster situations and highlighted some of his research works in this field. The audience was very interested in the topic, since it is highly relevant to Malaysia, and their field of studies and research. The program ended with a networking tea session while the audience took the opportunity to interact with the speaker. Prof. Jamalipour then travelled to UTM Kuala Lumpur to visit the UTM Ericsson Innovation Centre for 5G (IC5G) in the afternoon.

IEEE Vehicular Technology Society
Connecting the Mobile World
IEEE

Prof. Jamalipour during his lecture at UTM Skudai.

Prof. Jamalipour with some of the audience at UiTM Shah Alam.

Visit to UTM Ericsson Innovation Centre for 5G (IC5G).

On the following day, April 12, 2017, Prof. Jamalipour delivered his second lecture at MCMC, Cyberjaya. This lecture was held at the MCMC auditorium and co-organized with the Malaysian Technical Standards Forum Bhd (MTFSB) and MCMC. The lecture received overwhelming response and attracted more than 200 attendees. Prof. Jamalipour started his lecture by introducing IEEE, ComSoc and VTS to the attendees. He then elaborated on two topics, namely "Current and Future Research Challenges in Smart Grid Networks and the Internet of Things" and "Wireless Communications Techniques for Emergency and Disaster Situations." The lecture concluded with highlights on open research areas that generated a lot of interest among the attendees. The tutorial-style lecture helped attendees understand the content easily. Moreover, there was ample opportunity for question and answer after the lecture. The co-organizers were also generous to sponsor lunch for the speaker and all the attendees, where they took the opportunity for informal discussion and networking. The event ended around 1:45 pm, and Prof. Jamalipour went dirctly to the Kuala Lumpur International Airport for a flight to Johor Bahru (JB). In JB, he was welcomed by Chee Yen Leow, one of the ComSoc/VTS executive committee members, and later brought to The Pulai Springs Resort, Johor Bahru.

Prof. Jamalipour presented his third lecture on April 13, 2017, at the Faculty of Electrical Engineering, UTM Skudai, Johor. This event was co-organized with the Wireless Communication Centre

(Continued on Newsletter page 3)

IEEE ComSoc Lebanon Chapter: Winner of the EMEA 2017 Chapter Achievement Award

By Dr. Sarah Abou-Chakra, IEEE ComSoc Lebanon Chapter Chair

The Executive Committee of the ComSon Lebanon Chapter would like to convey their sincere thanks to the Regional Directors of IEEE ComSoc for considering the Lebanon Chapter as one of the winners for CAA–2017. We also would like to thank the past Executive Committee and in particular the past Chair, Professor Bachar El-Hassan, for initiating our 2016 achievements.

In 2016, the Chapter actively participated in several events in Lebanon, as detailed below.

On June 2, the Chapter technically sponsored a workshop held at the Doctoral school of the Lebanese University (EDST)–Tripoli. The workshop was on the topic "New Generation of Networks: Cloud Computing in 5G Networks." The event included four sessions: the opening session, Introduction to 5G, Cloud Computing for the Internet of Things, and 5G Security and Enterprise Vision. Eight speakers participated in this workshop and about 80 persons attended this event.

The Chapter technically sponsored and organized the Radio-Frequency and Microwave tracks in the Middle East Conference for Antennas and Propagation (MECAP 2016) that was held in Beirut from September 20 to 22, 2016. More than 60 researchers and Ph.D. students attended the conference.

We participated in the IEEE Day held on October 8, 2016 where Dr. Sarah Abou-Chakra, who was the chapter Vice-Chair, presented the latest chapter activities. In addition, members of the Executive Committee participated in the jury evaluating the student branches. More than 160 IEEE members joined the event at the Lebanese International University (LIU)–Jdeideh.

The Chapter also technically sponsored and organized the communications tracks at the IEEE International Multidisciplinary Conference on Engineering Technology (IMCET 2016) that was held in Beirut on November 2 and 3, and which attarcted 70 researchers.

The Chapter also organized several communications related activities, as detailed below.

The activities started with two Distinguished Lecturer Tours. In February, Prof. Mohamed-Slim Alouini from KAUST University, KSA presented two lectures on "5G, Evolution or Revolution?" These lectures attracted 70 attendees. Then in April, Prof. Andrei Gurtov gave two presentations on the topic "Software Defined Mobile Networks: Beyond LTE Architecture." These presentations were attende by 80 professors and students.

On May 7, the Chapter organized the fourth Lebanon Communications Research Day (IEEE LCRD'16) at the Doctoral School of Lebanese University (EDST)–Hadath Campus. The goal of this event is to allow Lebanese researchers and engineers working in the area of communications and networking to meet, present their work and exchange research ideas. Three invited speakers and 26 researchers contributed to IEEE LCRD'16. Sessions covered wireless sensor networks, future trends in wireless networks, network security, and cloud computing and networking. About 135 professors and students attended.

After five successful editions of IEEE LCRD, it will be upgraded to the Middle East and North Africa COMMunications Conference (MENACOMM'18) at the Holy Spirit University of Kaslik (USEK) in Jounieh, Lebanon on April 18–20, 2018 (www.menacomm-conference.com). The objective of MENACOMM is to bring researchers from academic institutions and industry from all over the world to the shores of the Mediterranean.

The sixth Lebanon Communications Student Competition (IEEE LCSC'16) was held on June 4, 2016, at AUST, Zahle Campus. The competition attracted 23 communications related projects from 12 Universities in Lebanon. The projects were divided into three groups: Algorithms and Simulations, Mobile Applications and Smart Phones and Systems, and Devices and Hardware. The evaluation process involved 15 professors from different universities and telecom companies. Three winners were selected from each group and an amount of 3000 USD, offered by Electricity of Zahle, was distributed to the winners with certificates from the IEEE ComSoc Lebanon Chapter. Attendees numbered about 150.

To commemorate the MOU signing between the IEEE Lebanon Section and the Order of Engineers and Architects (OEA)–Tripoli, the IEEE ComSoc Lebanon Chapter organized a workshop on November 2 at the OEA on the topic "The Mobile Telecommunication Evolution in Lebanon and the World." Three speakers participated in this workshop: Dr. Rola Naja, associate professor at the Lebanese University spoke about "Beyond 4G Networks: Evolution or Revolution?"; Rami Assoum, Head of the VAS Department at Alfa, Lebanon, spoke on the topic "An Insight on the Future of Mobile Business"; and Dr. Charbel Fares, professor at USEK, discussed the "Readiness of the Lebanese Telecom Infrastructure to integrate the Internet of Things." About 80 engineers from the OEA–Tripoli and IEEE members attended this workshop.

Last but not least, the Chapter celebrated the 10th Lebanon Communications Workshop (IEEE LCW'16), which has become the signature event of the ComSoc Lebanon Chapter and returned to its first location, the American University of Beirut (AUB), in November 2016. IEEE LCW'16 addressed "Tactile Internet"; extremely low latency in combination with high availability; reliability and security will define the character of the Tactile Internet. It will have a marked impact on business and society, introducing numerous new opportunities for emerging technology markets and the delivery of essential public services. We had a strong technical program with a highly distinguished group of speakers from Lebanon and abroad.

(Continued on Newsletter page 3)

Group photo of IEEE LCSC'16.

CHAPTER REPORT

Workshop on Communications, Signals and Information Processing (ComSIP)

By Cesar Vargas-Rosales and Rafaela Villalpando-Hernandez, Tecnológico de Monterrey, Mexico

The Monterrey Chapter of the IEEE Communications Society organized the workshop on "Communications, Signals and Information Processing (ComSIP)" on April 28, 2017 at Tecnologico de Monterrey in the city of Monterrey, Mexico. The main objective of the workshop was to gather academia and industry in one place to establish collaboration for future projects in the areas of ComSIP.

The first workshop on communications, signals and information processing put together 16 different presentations with research on the broad areas of signals, intelligent systems, information processing, and communications. There were presentations on topics such as signal propagation, vehicular communications, machine learning and its applications, sensor networks, digital signal processing, MEMS/NEMS, meta-heuristics, nanophotonics and adaptive algorithms. There was also a session with 15 different posters where everyone had the opportunity to discuss and looking into more topics with different authors. Some posters presented solutions to problems on MIMO systems, anomaly detection in networks, position location in sensor networks, ray launching for propagation prediction, silicon ion-channel sensors and mobile health monitoring among others. The opportunities to get involved in the research activities of ComSIP and these dynamic research areas were also introduced by interaction sessions between academia and industry.

More than 60 attendees included people from industry, undergraduate, graduate students and professors at Tecnologico de Monterrey. All attendees had the opportunity to listen to experts talk in different sessions and were able to explore different research areas in an intensive way.

International researchers that were invited as special guests in the workshop included Dr. Andreas Spanias (SenSIP, Arizona State University), Dr. Jennifer Blain Christen (SenSIP, Arizona State University), Dr. Emma Hart (Edinburgh Napier University, Scotland), and Dr. Gregory J. Gbur (University of North Carolina). Also among the attendees to the workshop were people from Axtel, AT&T and NIC México, which are telecommunications companies with a base in Monterrey.

The main room for the Workshop.

Organizers, volunteers and guests: Mahdi Zareei, Leyre Azpilicueta, Fausto Granda, Jennifer Christen, Rafaela Villalpando, Andreas Spanias, Jaime Zuñiga, and Cesar Vargas.

Other groups were also involved in the organization of the event such as the research focus group on Telecommunications and Networks from the School of Engineering and Sciences at Tecnológico de Monterrey, and Consejo Nacional de Ciencia y Tecnología (CONACYT) with financial support through the NSF/IUCRC-CONACYT/CoBI ComSIP project number 274733 where Dr. Cesar Vargas is the leader. The project ComSIP collaborates with Arizona State University through the SenSIP center led by Dr. Andreas Spanias.

Among sponsors, volunteers and organizers, we acknowledge CONACYT, NSF, the IEEE Monterrey Section, the IEEE ComSoc Monterrey Chapter, the IEEE-Eta Kappa Nu Lambda-Rho Chapter at Tecnologico de Monterrey and the SPIE Student Chapter.

GCN GLOBAL COMMUNICATIONS NEWSLETTER

STEFANO BREGNI
Editor-in-Chief
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

FABRIZIO GRANELLI
Associate Editor-in-Chief
University of Trento
Email: fabrizio.granelli@unitn.it

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

REGIONAL CORRESPONDENTS WHO CONTRIBUTED TO THIS ISSUE
EWELL TAN, SINGAPORE (ewell.tan@ieee.org)

IEEE ComSoc™
IEEE Communications Society

www.comsoc.org/gcn
ISSN 2374-1082

DISTINGUISHED LECTURER TOUR/*Continued from page 1*

(WCC) UTM. Here, Prof. Jamalipour covered the topic of "Current and Future Research Challenges in Smart Grid Networks and the Internet of Things." The event attracted 35 participants. In this lecture, the topic of Smart Grid networks was explained and the main challenges in such a network, including the security and privacy of grid customers, were highlighted. The lecture also included the new research and development direction toward the Internet of Things where the network connectivity extends its boundary beyond the usual digital devices to basically every device, appliance and "thing" around humans.

DLTs in Malaysia continues to be very successful and they will likely to be well attended in the future by academics, researchers and industry practitioners. Hosting such DLT events in various locations around Malaysia is an excellent way to encourage IEEE members and the technical community at large to participate in IEEE Comsoc/VTS technical dissemination programs. This surely helps improve recognition and generate growth for the IEEE ComSoc and VTS in Malaysia. The IEEE Malaysia ComSoc/VTS Joint Chapter would like to thank our speaker, Prof. Jamalipour, and our co-organizers and sponsors, UiTM Shah Alam, MTFSB, MCMC, and WCC UTM Skudai, for making this DLT a success. Above all, we would like to extend our special appreciation to IEEE ComSoc and IEEE VTS for arranging such a wonderful program.

LEBANON CHAPTER/*Continued from page 2*

In parallel with IEEE LCW'16, the chapter co-organized with the IEEE Young Professional Affinity Group a STEP event that included the signing of three MOUs with three renowned telecommunications' companies in Lebanon.

The chapter is trying, with its limited resources, to organize fruitful events in all parts the country while encouraging different students to become society members and enhancing industry-academia collaborations.

MILITARY COMMUNICATIONS

Kevin S. Chan

Frank T. Johnsen

Information sharing is the key concept of network-centric military operations. As the complexity of operational environments increases, military communications networks must be agile to meet strict and evolving mission requirements in the midst of challenges presented by dynamic operational environments. The RF spectrum, for example, is a limited resource, which can be utilized better and in a more flexible way through the use of software defined networking and cognitive radios. Tactical communications encompasses all communication technologies used in the battlefield, constituting a heterogeneous collection of approaches and technologies. The battlefield is typically a disruptive environment, often called DIL (short for disconnected, intermittent, and limited), which summarizes the challenges that engineers and tactical network operators constantly face. Thus, it is important that capabilities such as self-organization, decentralization, and delay tolerance are implemented to facilitate information sharing in such environments. To assist in timely and accurate decision making, rapid information sharing is an important aspect of information superiority. Machine learning techniques can be used to analyze large streams of data, but military communications must ensure the security of the data and techniques. If automated sharing across different networks and security levels can be provided with adequate levels of trust and data leak prevention, one important building block in cross-domain information sharing will be in place. The four articles that constitute this year's Feature Topic on Military Communications provide an overview of these trends in military communications.

The first article, ″Unified Solution to Cognitive Radio Programming, Test, and Evaluation for Tactical Communications,“ presents challenges related to spectrum limitations, and proposes a solution to software-defined radio programming. The design and implementation of a radio software programming tool is presented, along with an approach to making prototyping of cognitive radio capabilities with tactical radios in a fast and cost-effective manner.

The article "Combining Software-Defined and Delay-Tolerant Approaches in Last-Mile Tactical Edge Networking" addresses interconnecting higher-level networks with abundant resources with the tactical edge. The authors identify the last mile as the major challenge, and suggest combining software defined and delay-tolerant approaches to better support applications and information sharing in such networks.

In addition to delay tolerance, there are other aspects that can improve tactical communications as well. The article "Inband Full-Duplex Radio Transceivers: A Paradigm Shift in Tactical Communications and Electronic Warfare?" discusses one such aspect. Here, it is pointed out that inband full-duplex operation has great potential in civilian wireless communication, since it has the potential to double transmission links' spectral efficiency. The authors aim to exploit this emerging radio technology in military communication applications. The key concept is STAR, short for simultaneous transmission and reception, which allows a radio to do more than one thing at the same time, for example, allowing radios to conduct electronic warfare at the same time as they are receiving or transmitting other signals.

The final article, "Data Leakage Prevention for Secure Cross-Domain Information Exchange,“ shows how a trusted component, a data guard, can be leveraged to achieve secure cross-domain information exchange. However, the authors point out that while a guard is a high-assurance component, a weakness of the guard-based solution is that there is often limited assurance in the correctness of the security labels. Hence, the authors propose to use advanced content checking as an additional measure to prevent data leaks. Leveraging machine learning techniques enables data-driven methods that automatically infer the words associated with classified content. The article discusses a proof-of-concept implementation and application of this approach.

The Editors are confident that these articles provide our readers with some of the current developments and challenges in the military application of communications. The four articles that make up this year's Feature Topic were selected from a total of 13 high-quality manuscripts received. We encourage developers and researchers working on military communications and related fields to consider submitting a paper to the Feature Topic on military communication in 2018.

BIOGRAPHIES

KEVIN CHAN [M] (kevin.s.chan.civ@mail.mil) is a research scientist with the Computational and Information Sciences Directorate at the U.S. Army Research Laboratory (ARL), Adelphi, Maryland. His research interests are in network science and dynamic distributed computing. He is involved in ARL's collaborative programs in network science and distributed analytics. He has been a member of several NATO panels to study C2 Agility and networked services for tactical networks. He holds Ph.D. and M.S. degrees in electrical and computer engineering from Georgia Institute of Technology, Atlanta, and a B.S. in ECE/EPP from Carnegie Mellon University, Pittsburgh, Pennsylvania.

FRANK T. JOHNSEN (Frank-Trethan.Johnsen@ffi.no) is a principal scientist at the Norwegian Defence Research Establishment (FFI). He is currently working within the area of information and integration services, with a special focus on applying service-oriented architecture (SOA)in the tactical domain. He also holds a part-time position as associate professor at the University of Oslo, which involves supervising students and teaching SOA. He received his Cand.scient. and Ph.D. degrees from the University of Oslo.

A Unified Solution to Cognitive Radio Programming, Test and Evaluation for Tactical Communications

Yalin Sagduyu, Sohraab Soltani, Tugba Erpek, Yi Shi, and Jason Li

The authors present the design and implementation of a visual and modular radio software programming tool that supports easy, fast, and radio-agnostic development of cognitive radio and network protocols and security mechanisms. This tool is fully integrated with a unified T&E framework that applies the same SDR solution to high fidelity simulation and emulation tests under a common, controllable, and repeatable scenario.

Abstract

Spectrum is limited, but the demand for it is growing steadily with new users, applications, and services in both commercial and tactical communications. The current paradigm of static spectrum allocation cannot satisfy this demand, resulting in a congested, contested environment with poor spectrum efficiency. Tactical radios need to share spectrum with other in- and out-of-network tactical and commercial radios, subject to potential jamming and other security attacks. Cognitive radio provides tactical communications with new means of spectrum sharing, cohabitation configurability, and adaptation to improve communication rates, connectivity, robustness, and situational awareness, all translated to network-centric mission success. A systematic solution to SDR programming and test and evaluation (T&E) is needed to address spectrum challenges with the network-centric mission success set as the primary goal. Once equipped with cognitive radio capabilities, tactical radios can quickly and reliably discover the white-space and effectively use spectrum opportunities across time, space, and frequency. This article presents the design and implementation of a visual and modular radio software programming tool that supports easy, fast, and radio-agnostic development of cognitive radio and network protocols and security mechanisms. This tool is fully integrated with a unified T&E framework that applies the same SDR solution to high fidelity simulation and emulation tests under a common, controllable, and repeatable scenario. This unified approach makes prototyping of cognitive radio capabilities with tactical radios faster, easier, and cost-effective.

Introduction

Spectrum resources are scarce, and they are not efficiently utilized with a static and dedicated allocation of frequency bands and their sporadic use. Tactical communications are typically assigned legacy frequency bands for exclusive spectrum use. However, spectrum becomes crowded with the growing demand (e.g., multimedia) of commercial communications, for example, LTE, WiFi and unlicensed national information infrastructure (U-NII) devices. Therefore, new means are necessary to enable tactical radios to share spectrum with commercial users (e.g., mobile and fixed wireless broadband use, LTE, and electronic news gathering). For instance, AWS-3 auction aims to release 1695–1710 MHz, 1755–1780 MHz, and 2155–2180 MHz frequency bands, raising both opportunities and challenges for tactical and commercial users to share the spectrum.

Tactical networks span a significant part of available spectrum resources, and should be designed, implemented, tested, and evaluated with the goal of improving spectrum efficiency and mission success. The legacy form of a static spectrum allocation poses a major obstacle for efficient use of limited wireless resources. This challenge has promoted substantial research and development into cognitive radio and software-defined radio (SDR) technologies with network-level perception, learning, adaptation, and optimization for efficient spectrum utilization. Cognitive radio emerges as the enabling technology to support configurability of radios on the fly and access to broader pools of spectrum, and more efficient utilization of current wireless resources, playing a key role for the next generation of both commercial and tactical communications [1].

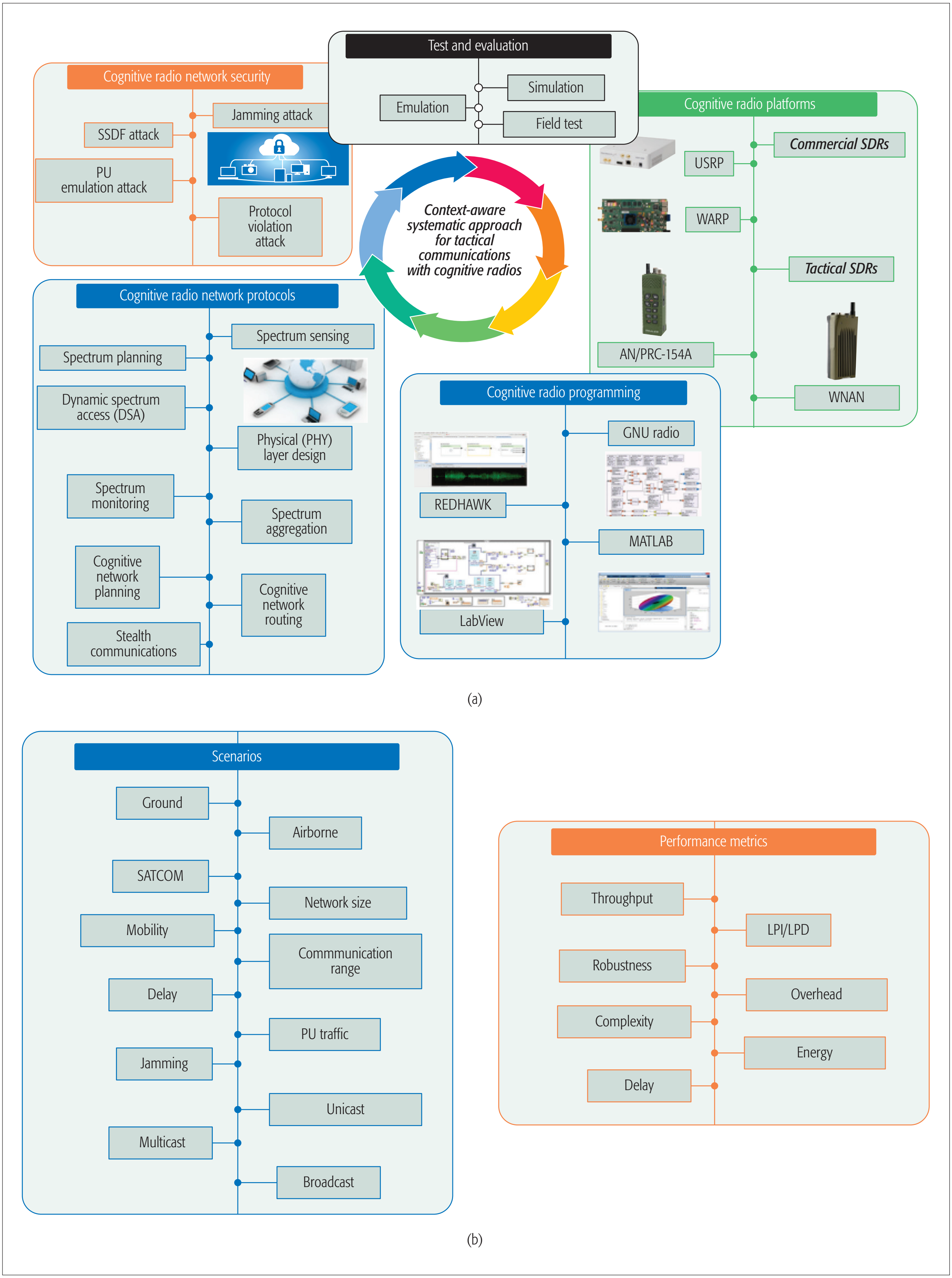
To build the next generation of tactical network communication systems with cognitive radios, a systematic approach including a *full scope of design, implementation, and test and evaluation (T&E) environment* is necessary with the systematic and unified execution of the following five components (Fig. 1a).

Network Protocols: Cognitive radio uses a "cognitive engine" to learn the spectrum accurately with small overhead, and then promptly adapts to spectrum dynamics, including channel occupancy, interference, congestion, and mobility. Instead of a separate and static protocol abstraction, cross-layer design appears a viable way to integrate spectrum sensing and dynamic spectrum access (DSA), routing, and other network protocols in a unified cognitive network protocol stack. This protocol stack includes physical (PHY), medium access control (MAC), and higher layers that optimally strike a balance between performance and overhead/complexity trade-offs.

Network Security: The configuration and adaptation of cognitive radio is vulnerable to various forms of attacks due to the broadcast nature of the wireless medium in anti-access/area-denial

Digital Object Identifier:
10.1109/MCOM.2017.1700222

The authors are with Intelligent Automation, Inc.



(a)

(b)

Figure 1. a) Context-aware systematic approach for tactical communications with cognitive radios; b) scenarios and performance metrics for tactical networks.

The goal is to improve network-centric mission success according to various criteria including but not limited to assured connectivity, coexistence (cohabitation), spectrum situational awareness, minimum latency, maximum geographic, bandwidth and time availability of spectrum resources and cyber protection.

(A2AD) environments. These attacks range from jamming to insider threats including spectrum sensing data falsification (SSDF), primary user (PU) emulation, and protocol violation attacks. Novel means are sought to combat these security issues and maintain the mission-critical performance in tactical networks of cognitive radios.

Platforms: The use of commercial radios is likely the first step to quickly prototype, test, and evaluate cognitive radio capabilities. The next step is to implement these capabilities with tactical radios. Once mature, cognitive radio technologies can be ported to tactical SDRs with different levels of size, weight, and power-cost (SWaP-C) characteristics and flexibility to enable cognitive radio capabilities.

Programming: Cognitive radios require fast and efficient SDR programming of protocol modules. Various commercial tools (e.g., GNU Radio, REDHAWK, LabView, and MATLAB) are available to program and control SDRs at the software level or field programmable gate array (FPGA) level for general signal processing purposes typically focused on the PHY layer. In addition, higher-layer protocols are needed as part of SDR programming. For fast implementation, there is a growing demand for a user-friendly visual tool that can automatically generate radio-agnostic codes for the full protocol network stack based on specified modules in the SDR architecture.

Test and Evaluation: A cognitive radio involves various tunable parameters to be extensively tested for seamless integration of network protocols. The standard procedure is to separate simulation studies, such as Common Open Research Emulator (CORE) [2] and Extendable Mobile Ad Hoc Network Emulator (EMANE) [3] and emulation tests with radios. A systematic T&E approach is needed to unify them under a common, controllable, and repeatable scenario: a) scalable and high fidelity simulations and b) in-lab radio-in-the-loop emulation tests with controlled RF characteristics.

The goal is to improve network-centric mission success according to various criteria including but not limited to assured connectivity, coexistence (cohabitation), spectrum situational awareness, minimum latency, maximum geographic, bandwidth, and time availability of spectrum resources and cyber protection. Potential scenarios and performance metrics are shown in Fig. 1b. To achieve this goal, this article presents the novel design and implementation of a *visual and modular* radio software programming tool to support *easy*, *fast*, and *radio-agnostic* development of cognitive network protocols and security mechanisms. This fully developed capability is integrated with a *unified T&E environment* that applies the same SDR solution to *high fidelity simulation and emulation tests* under the common scenario control. The novelty of this unified solution consists of two main components.

EZPro: A visual and modular radio software programming tool is presented to generate a workflow of cognitive radio protocols and security mechanisms, develop individual modules for cognitive radio functionalities, and automatically generate the end-to-end software code that can be used with simulation and emulation tests (with real radios).

Network Control Environment: A unified network control environment is presented to set up nodes, configure them as virtual (for simulation tests) or real (for emulation tests), set up the network scenario (including topology, channel, and mobility effects), assign the SDR code to virtual nodes in simulations, or port the SDR code to SDR platforms for emulation tests.

- ***Simulation Environment:*** High fidelity simulation is based on CORE and EMANE. CORE provides the high fidelity simulation environment with a real network stack and traffic for emulating networks on one or more machines [2]. PHY/MAC is emulated by various models (e.g., RF pipe or 802.11) of EMANE [3].
- ***Network Channel Emulator:*** The same SDR code used for simulations can be used for emulations tests with real SDR platforms. Network channel emulators are needed to emulate wireless channels among SDRs in in-lab tests and support controllable, repeatable, and scalable hardware-in-the-loop experiments. In the featured setup, a network channel emulator (a radio channel emulator supporting a full mesh network connectivity), RFnest [4], is used to emulate radio channels and RF interactions among radios, and support potential interactions of real (emulated) and virtual (simulated) radios. RFnest adjusts channel conditions (e.g., attenuation, delay, multipath, and Doppler) according to the scenario generated for tactical network communications.

With this unified SDR programming and T&E environment, a user can easily, quickly, and reliably design and implement cognitive radio protocols across the network protocol stack, and test and evaluate them with both simulation and in-lab hardware-in-the-loop emulation tests under realistic tactical network scenarios.

The rest of the article is organized as follows. The following section presents a novel systematic solution to network-centric SDR programming. Then we present a systematic unified solution to cognitive radio network T&E. The final section concludes the article.

A Novel Systematic Solution to Network-Centric SDR Programming

State-of-the-Art SDR Programming Tools

There are various ways to program cognitive radios. Some examples of the state-of-the-art SDR programming software are listed below:

GNU Radio: GNU Radio [5] is a free and open source software development toolkit that provides signal processing blocks to implement software radios. GNU Radio can be used with existing SDRs, such as universal software radio peripherals (USRPs), or without hardware in a simulation-like environment.

REDHAWK: REDHAWK [6] is an SDR framework to support real-time software radio applications. REDHAWK develops and tests software modules called "Components" and composition of Components into "Waveform Applications" deployed on a single or multiple network-enabled computer(s).

LabView: LabVIEW Communications [7] from National Instruments offers a graphical design environment integrated with the USRP. LabVIEW offers parallel constructs spanning multiple cores, threads, and targets, and supports real-time embedded processors and FPGAs.

MATLAB: MATLAB and Simulink [8] are used for wireless design, simulation, and analysis on SDRs such as setting up SDR hardware with preconfigured radio functions, and performing real-time signal analysis and measurement. MATLAB and Simulink provide support packages for popular SDR hardware such as USRP, Zynq [9], and RTL-SDR [10] Radio.

These existing SDR programming tools are typically focused on digital signal processing (DSP) at the PHY layer. The extended capabilities that are needed to implement cognitive radio solutions effectively with tactical radios include:

- Full network protocol stack (beyond PHY/DSP)
- Executable, radio-agnostic code
- The same code used for simulation and hardware testing
- Operating system independence
- Dynamic modular workflow design
- User-friendly GUI with visual programming tools
- Modular I/O with message passing
- Visual module manager with edit/merge functions

EZPro as a Visual and Modular SDR Programming Tool

To address these gaps, this article presents the design and implementation of EZPro as a visual and modular "Easy Programming" environment to generate SDR software. The user of this software either leverages the built-in code blocks or generates new code blocks. A user designs a workflow by dragging different blocks into the design panel and connecting them, and then tests the workflow in the same environment. EZPro then generates standalone software from the workflow that is portable and executable outside of the EZPro environment.

A multithreaded back-end Python class with input/output ports and input arguments supports each block. Different blocks communicate with each other through their connections over a queueing system that supports multi-threaded and multi-processor communication. EZPro provides user tools to generate software for various network layers. Any existing software (e.g., GNU radio software blocks) can also be wrapped as an EZPro block and reused in the workflow.

Figure 2 shows an example EZPro workflow for the SDR software that will run at each cognitive radio. This workflow consists of several blocks for different cognitive radio functionalities:

- The *spectrum sensing* block senses the environment and feeds the spectrum occupancy results to the *attack countermeasures* and *DSA blocks*. Additionally, a *spectrum database* block can also be implemented that connects to a database and pulls the expected spectrum occupancy results.
- The *DSA* block evaluates the incoming information regarding the spectrum occupancy and determines the operating frequency.
- The *routing* block decides on the next hop for the packets.
- The *attack countermeasures* block analyzes the spectrum sensing results and decides what mitigation techniques will be used to ensure continuous operation.

The user can visually revise/extend the workflow. The four main steps in executing EZPro (Fig. 3) are:

1. Drag an individual code block from existing blocks or design an individual block to make the workflow (EZPro provides designated spots in the template to insert the user code).
2. Connect blocks in the workflow by using the Tip tool (there is a separate enumerated connection for each parameter passed among blocks).
3. Configure each block using the Parameter Panel.
4. Automatically generate source code for the workflow (use EZPro interactive tools to run and debug program source code) and execute the program.

After these four steps, program source code is automatically converted to standalone software that is deployable to any systems with general-purpose processors (GPPs).

The computational cost of running EZPro software alone is negligible (about 0.3 percent CPU and 2.9 percent memory usage measured with a laptop equipped with Intel Core i7 — 2.9 GHz CPU and 8 GB RAM). The computational cost of running the developed software depends on the implementation complexity, such as the number of threads and the number of blocks used.

EZPro aims to make the programming of tactical SDR and prototyping cognitive radio capabilities faster, easier, and cost-effective. The SDR software generated through EZPro is radio-agnostic and can run on any GPP. This way, we can seamlessly integrate cognitive radio capabilities into tactical radios without any change in radio architecture. Note that once the radio code is generated by EZPro and deployed at a radio, it runs without EZPro installation. The same code can be used with commercial SDRs for fast prototyping and evaluation of cognitive radio network technologies before implementation on tactical SDRs.

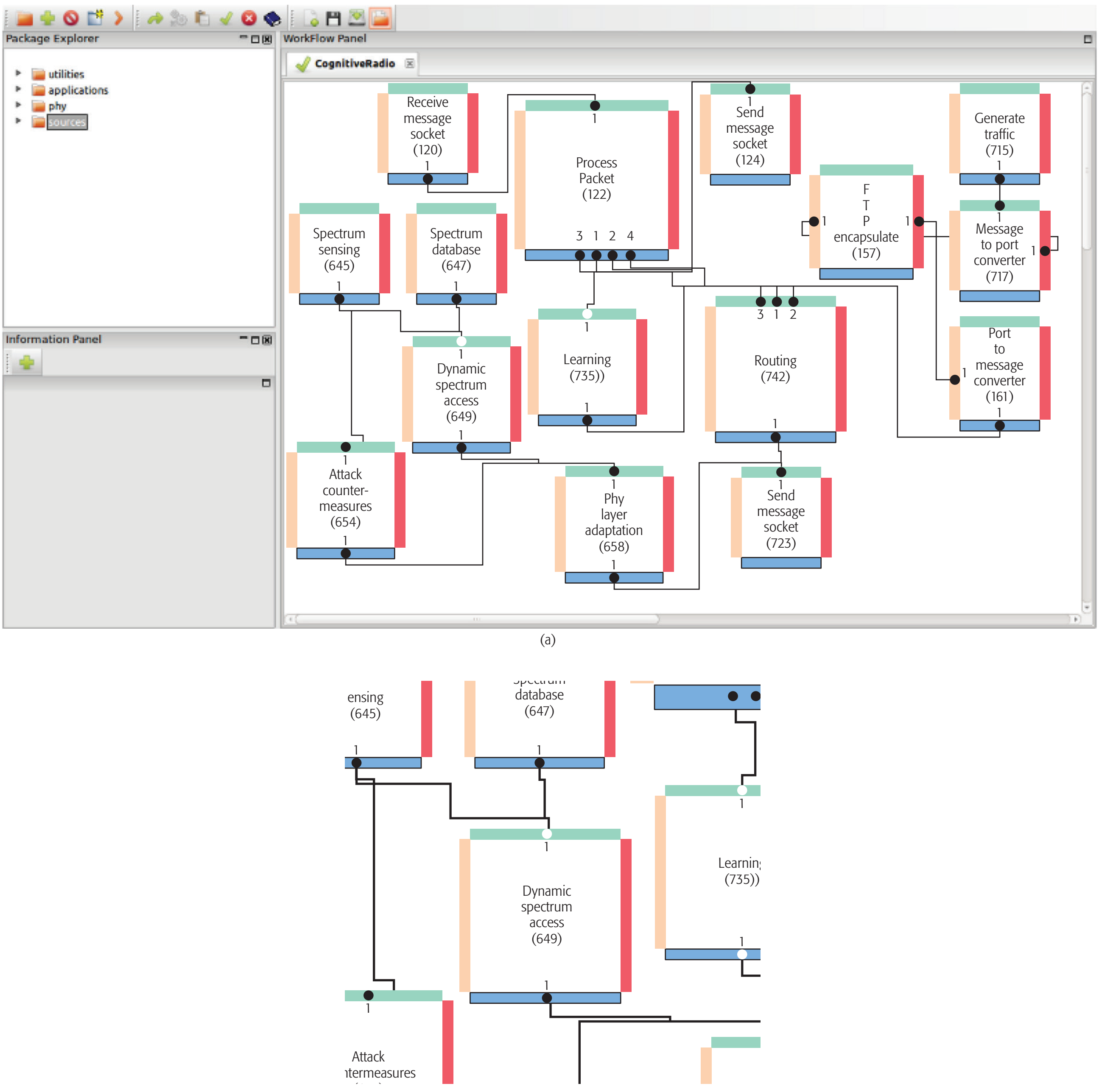
Popular examples of commercial SDRs that can be programmed with EZPro are USRPs Wireless Open Access Research Platform, (WARP), bladeRF, HackRF, and Matchstiq. Examples of tactical SDRs are AN/PRC-154 (Rifleman Radio), AN/PRC-152, and WNaN, which can be programmed using a peripheral controller or through implementation on a radio processor.

SDR Modules for EZPro Implementation

Cognitive radio is driven by a cognitive engine that integrates various network protocols, ideally both adaptive and robust. EZPro has been used for spectrum sensing, DSA, routing, and security functionalities for cognitive radios. Below, we give a summary of cognitive radio network functionalities that can be implemented as modules in the presented SDR programming environment.

Spectrum Sensing: Cognitive radios discover spectrum opportunities in terms of idle channels via spectrum sensing [11]. A channel can

The user of this software either leverages the built-in code blocks or generates new code blocks. A user designs a workflow by dragging different blocks into the design panel and connecting them, and then tests the workflow in the same environment. EZPro then generates standalone software from the workflow that is portable and executable outside of the EZPro environment.



(a)

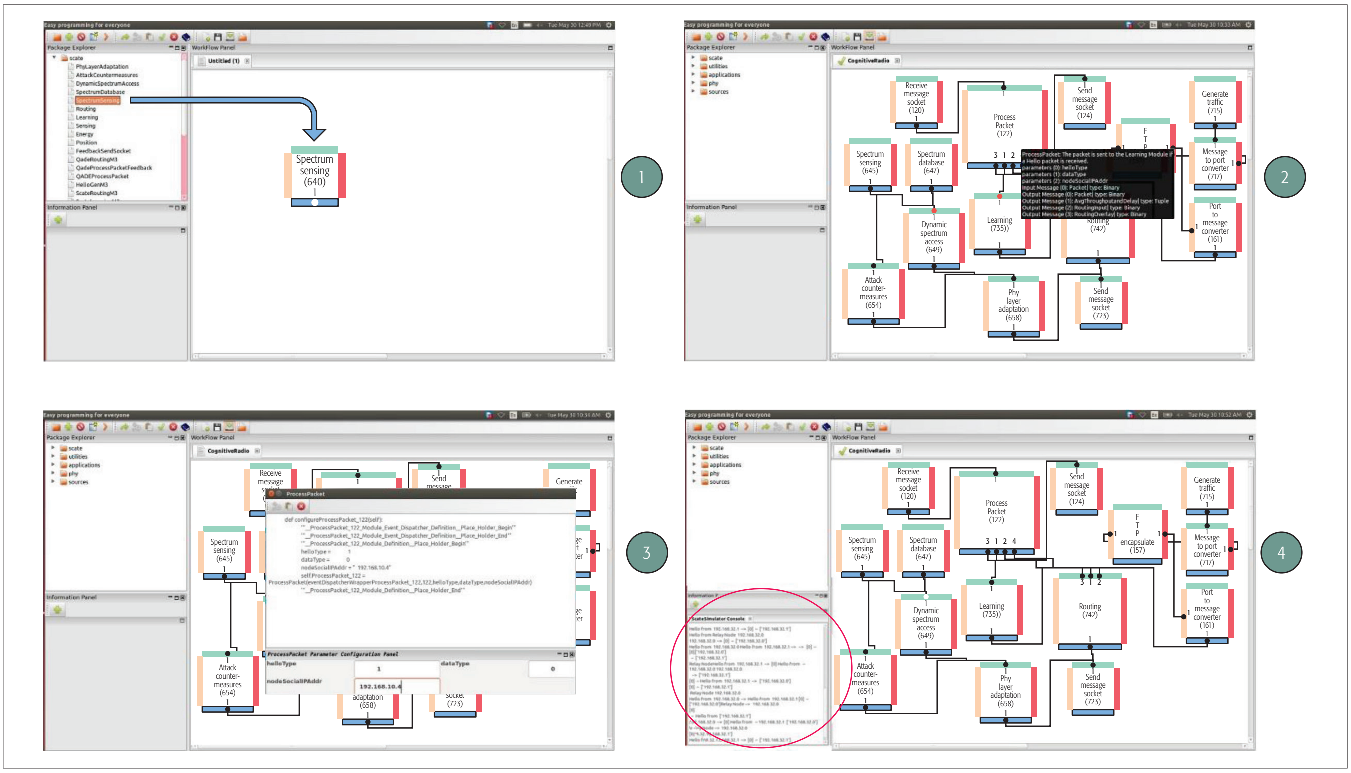
(b)

Figure 2. a) Example of EZPro workflow for SDR software; b) a closer look at an EZPro block with its connections to other blocks.

be detected as idle or busy by comparing the accumulated energy of received signal to a predetermined threshold (energy detector) or by exploiting the embedded features in cognitive radio signals (cyclostationary sensing). To prevent poor sensing sensitivity and high sensing error due to device-level energy constraints, poor signal quality (due to fading or mobility), and the hidden terminal problem, cooperative (or collaborative) sensing can combine measurements of multiple sensors into one common decision, either by soft combining of channel measurements or by hard combining of binary channel sensing results.

DSA: PUs (e.g., TV broadcast or radar) have dedicated channels, while tactical radios in the role of secondary users (SUs) need to access channels sensed to be available. Otherwise, if there are SUs only working on an unlicensed frequency, it is important not to interrupt existing transmissions. This hierarchical spectrum sharing is typically facilitated by interference avoidance (no interference to PUs and existing SU transmissions is allowed). Channel estimation and neighborhood discovery are two integral parts of DSA to support real-time spectrum monitoring. The DSA problem can also be translated to stealth (covert) communications, where tactical radios attempt to access the spectrum without being detected by other radios.

Cognitive Network Routing: In a multihop network communication setting, spectrum occupancy is time- and location-dependent. Therefore, joint design of routing and DSA is needed without



ProcessPacket: The packet is sent to the Learning Module if a Hello packet is received.
parameters (0): helloType
parameters (1): dataType
parameters (2): nodeSocialIPAddr
Input Message (0): Packet| type: Binary
Output Message (0): Packet| type: Binary
Output Message (1): AvgThroughputandDelay| type: Tuple
Output Message (2): RoutingInput| type: Binary
Output Message (3): RoutingOverlay| type: Binary

```
    def configureProcessPacket_122(self):
        '''__ProcessPacket_122_Module_Event_Dispatcher_Definition__Place_Holder_Begin'''
        '''__ProcessPacket_122_Module_Event_Dispatcher_Definition__Place_Holder_End'''
        '''__ProcessPacket_122_Module_Definition__Place_Holder_Begin'''
        helloType =        1
        dataType =         0
        nodeSocialIPAddr = " 192.168.10.4"
        self.ProcessPacket_122 =
ProcessPacket(eventDispatcherWrapperProcessPacket_122,122,helloType,dataType,nodeSocialIPAddr)
        '''__ProcessPacket_122_Module_Definition__Place_Holder_End'''
```

ProcessPacket Parameter Configuration Panel

helloType	1	dataType	0
nodeSocialIPAddr	192.168.10.4		

Figure 3. Four steps of EZPro execution.

maintaining any end-to-end path [12]. For example, the backpressure algorithm can be used to optimize a spectrum utility that consists of channel quality (e.g., RSSI) and congestion (e.g., differential queue backlog), and combine routing and channel access decisions in a cross-layer optimization framework.

Attack countermeasures: Defense mechanisms to a variety of attacks including jamming SSDF, PU emulation, and protocol violation attacks are needed.

•***Jamming Attack:*** In A2/AD environments, tactical radios may receive interference from in-network or out-of-network transmissions as well as deliberate jammers.

•***SSDF Attack:*** Cooperative sensing makes the cognitive radios vulnerable to SSDF attacks [13], where attackers may flip their sensing results before they report them to the fusion center with the objective of either blocking transmission initiatives or causing transmission failures.

•***PU Emulation Attack:*** The goal of the PU emulation attack is to create an intentional false positive for spectrum sensing by spoofing the role of a PU and generating a waveform similar to that of the PU.

•***Protocol Violation Attack:*** Insider threat arises when users start manipulating the underlying protocols and falsifying the information exchange. These attacks are typically low-signal and aim to degrade the performance slowly over time.

As a defense, cognitive radios can sense the spectrum and hop to available channels by tracking and adapting to malicious radio behavior [14] (e.g., by assigning and maintaining trust (or reputation) for each user, monitoring their activities and updating their trusts via consistency check with some type of feedback (e.g.,, RF signature or protocol behavior).

A Systematic Unified Solution to Cognitive Radio Network Test and Evaluation

Cognitive radio technologies applied to tactical communications should be tested and evaluated extensively before full-scale integration with military systems. Test and evaluation ranges from simulations to the use of over-the-air wireless testbeds and hardware-in-the-loop network emulation tests.

Simulations: There are various network simulators, such as ns-2/3, OPNET, and QualNet, that can be used to test and evaluate cognitive radio network protocols. They simulate traffic, protocol, and PHY effects. Other advanced simulators such as CORE and EMANE can be built by using real protocol stacks and generating real packet traffic that is carried among nodes represented by virtual machines.

Over-the-Air Wireless Testbed: Simulations use simplistic physical layer modeling that ignores various hardware effects, for example, nonlinearity, filtering, and intermodulation by hardware. There have been several developments of cognitive radio or SDR testbeds (e.g., ORBIT in WINLAB) [15]. These testbeds can represent static scenarios that cannot be reliably controlled or repeated and cannot emulate a representative geometry of a specific (e.g., mobile) scenario, as often needed to test the refined adaptation nature of cognitive radio functionalities.

Emulation Tests: Network channel emulators such as RFnest [4] fill this gap by providing a repeatable and controllable testbed environment for cognitive radios communicating with each other under dynamic RF propagation and mobility conditions that can be controlled and replayed.

Combining Simulation and Emulation Tests:

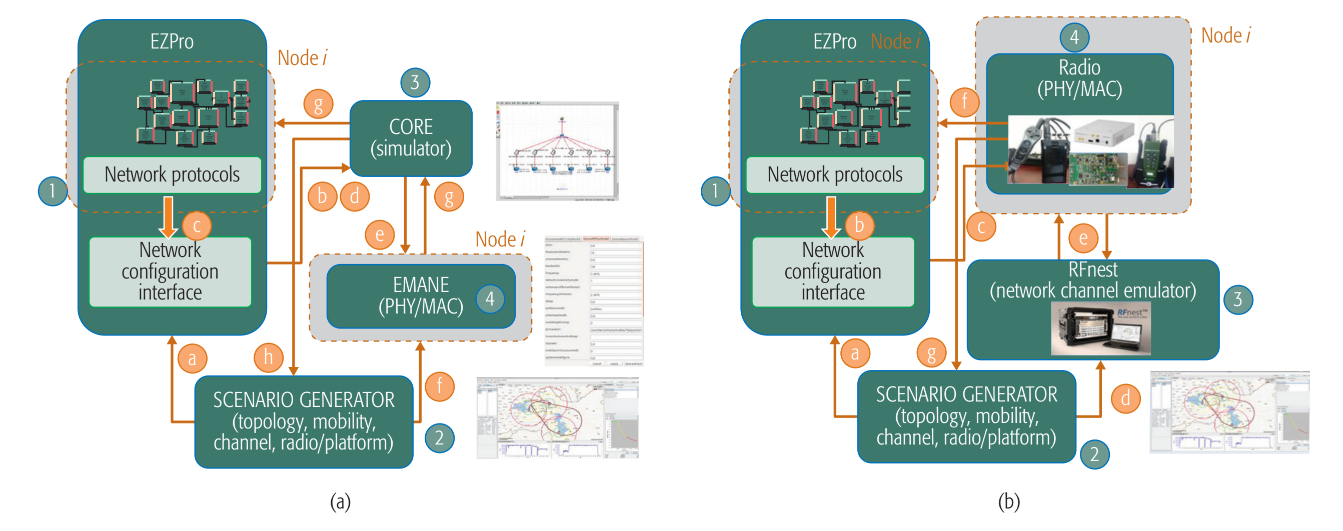

EZPro
Node *i*
Network protocols
Network configuration interface
CORE (simulator)
EMANE (PHY/MAC)
SCENARIO GENERATOR (topology, mobility, channel, radio/platform)
1
2
3
4
a
b
c
d
e
f
g
h
(a)
EZPro
Node *i*
Radio (PHY/MAC)
RFnest (network channel emulator)
Network protocols
Network configuration interface
SCENARIO GENERATOR (topology, mobility, channel, radio/platform)
(b)

Figure 4. A systematic unifying approach to combine testing: a) simulation; b) emulation.

We present a systematic unifying approach, illustrated in Fig. 4, to combine simulations and emulation tests, where simulation and emulations tests are controlled under a common scenario by using the same SDR code. In this setup, step 1 is to generate the SDR code (as discussed previously); step 2 is to generate and control the scenario; step 3 is to set up the interactions among cognitive network nodes (virtual CORE nodes in simulations and actual radios in emulations); and step 4 is to set up the wireless environment (EMANE for simulations and RFnest for emulations).

The following steps are followed to simulate a cognitive network solution (Fig. 4a):

a) The EZPro Network Configuration Interface imports the mission scenario from the Scenario Generator (available in RFview software) that generates the topology, mobility, channel, platform, and radio properties.

b) Using this network scenario, the Network Configuration Interface connects to CORE and automatically generates the CORE scenario with virtual nodes where each node represents a specific platform/radio. The Network Configuration Interface provides an automated tool to configure each CORE virtual node with standalone software provided by EZPro.

c) Each node in the network is initialized by EZPro with the traffic generator and network protocol stack wrapped as a standalone software executable. This software is identical to the one that would be ported to an actual radio and is run on the CORE simulator.

d) After this configuration stage, the Scenario Generator initiates the CORE simulator that runs virtual nodes with EZPro-generated standalone software.

e) In CORE simulation, EMANE provides the PHY/MAC for different waveforms, and higher-layer network protocols configure tunable knobs in waveforms available in EMANE. Default EMANE parameters can be configured manually by the user before running the simulation or automatically (according to higher layers) during the simulation.

f) The Scenario Generator sends channel conditions, mobility, and other information (through multicast packet updates) to EMANE. Accordingly, EMANE reconfigures its parameters in near real time (e.g., less than 1 ms delay) to simulate the wireless network environment.

g) While playing the mission scenario, each virtual node obtains feedback from the CORE/EMANE environment. This information is brought to higher network layers.

h) Performance results measured for each virtual node are sent to the Scenario Generator with a customized GUI to monitor performance (at either the network or node level).

To run a cognitive network solution on radios in emulation tests, steps a and b are the same as steps a and b in simulations, step g is the same as step h in simulations, and the following additional steps are followed (Fig. 4b):

c) After this configuration stage, the Network Configuration Interface ports network protocols to the radio.

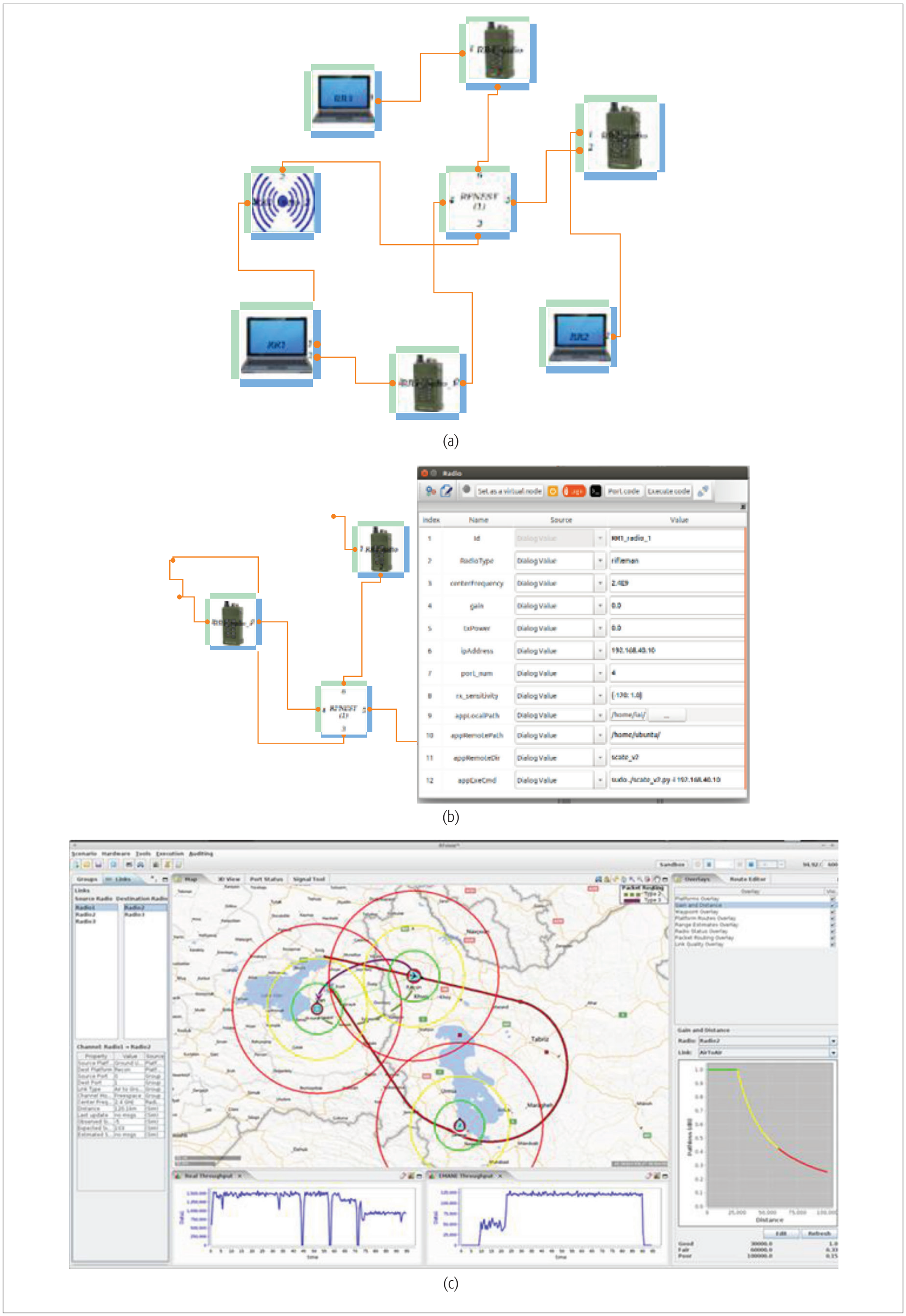
d) The Scenario Generator sends channel conditions, mobility, and other information (through multicast packet updates) to RFnest to emulate the wireless network environment.

e) Radio signals go through RFnest between radios, and RF channel conditions are adjusted.

f) While playing the mission scenario, each radio learns the RF environment (e.g., spectrum sensing), and this information is brought to higher network layers.

The network control GUI (Fig. 5a) implemented within EZPro is used to set up the testbed, which consists of RFnest, radios connected to the RFnest, and platforms attached to radios (e.g., the control elements that host the radio software to control the radio). This GUI allows the user to configure nodes (platforms) and radios as virtual or emulated (Fig. 5b), assign the SDR software with virtual nodes in CORE, or port it to radios.

The network control GUI is hosted on the control station. CORE is used as the network emulator, while EMANE provides the MAC and PHY layers in the virtual tests. The same radio software generated by the presented SDR programming approach (or through other means such as GNU Radio) is used for both virtual and emulation tests.



Index	Name	Source	Value
1	id	Dialog Value	RM1_radio_1
2	RadioType	Dialog Value	rifleman
3	centerFrequency	Dialog Value	2.4E9
4	gain	Dialog Value	0.0
5	txPower	Dialog Value	0.0
6	ipAddress	Dialog Value	192.168.40.10
7	port_num	Dialog Value	4
8	rx_sensitivity	Dialog Value	[-130: 1.0]
9	appLocalPath	Dialog Value	/home/lai/
10	appRemotePath	Dialog Value	/home/ubuntu/
11	appRemoteDir	Dialog Value	scate_v2
12	appExeCmd	Dialog Value	sudo ./scate_v2.py -i 192.168.40.10

Figure 5. a) Loading network scenario elements into EZPro; b) platform and radio configuration; c) RFview software to monitor performance.

The network control GUI is hosted on the control station. CORE is used as the network emulator, while EMANE provides the MAC and physical layers in the virtual tests. The same radio software generated by the presented SDR programming approach (or through other means such as GNU Radio) is used for both virtual and emulation tests.

Transmit power, frequency, signal bandwidth, antenna pattern, and gain are some of the configuration parameters for the radios. The EMANE model works at the packet level. Signal-to-interference-plus-noise ratio (SINR) vs. probability of reception curves are used to determine whether a packet is successfully received or not. IEEE 802.11a/b/g and time-division multiple access (TDMA) are example MAC protocols supported by EMANE software.

RFnest is used to emulate wireless channels. The antennas of the radios are removed, and the radios are plugged into the RFnest with RF cables. RFnest digitally controls channel properties between nodes. In this setup, radios do not distinguish whether signals are coming over the air or through RF cables. RFnest can replay channel (I&Q) data and support virtual nodes and interactions with real radios.

RFview software is used to define the test scenarios (Fig. 5c) including topology, mobility, and channel effects. The number and position of the radios are specified on the map. Mobility pattern can be either provided in advance or updated on the fly for each radio. The PHY layer parameters such as the transmit power and frequency are specified for each radio before starting the tests. RFnest includes various path loss models. The path loss model to be used during the tests is also selected using the RFview GUI. The user can create charts on the RFview GUI to monitor the network's performance during runtime. The packet exchanges between the radios can also be visualized on the map during the tests.

The solution presented in this article provides fast, reliable and cost-effective prototyping of context-aware systematic design along with realistic and relevant experimentation as essential means to achieve the performance gains offered by cognitive radios for network-centric mission success.

Field Tests

After the initial T&E of cognitive radio technologies demonstrates the feasibility, the next step is to conduct field tests with tactical radios in realistic and relevant environments before full-scale development and integration. The technology readiness level increases with each step of T&E, and the goal at each step is to demonstrate the improvement of spectrum efficiency and mission success. The unified T&E environment for simulations and emulation tests aims to reduce the time from protocol development to field test, and increase the success of a field test by evaluating various potential scenarios and identifying problems before the field test.

Conclusion

Cognitive radios provide a new communication paradigm to increase spectrum efficiency. This article addresses spectrum challenges for tactical communications and presents a systematic and unified SDR programming, test and evaluation solution with the necessary components to enable cognitive radio solutions for tactical communications. This solution provides fast, reliable, and cost-effective prototyping of context-aware systematic design along with realistic and relevant experimentation as essential means to achieve the performance gains offered by cognitive radios for network-centric mission success.

References

[1] B. Fette, "Fourteen Years of Cognitive Radio Development," *Proc. IEEE MILCOM*, San Diego, CA, Nov. 18–20, 2013.

[2] J. Ahrenholz *et al.*, "Core: A Real-Time Network Emulator," *Proc. IEEE MILCOM*, San Diego, CA, Nov. 16–19, 2008.

[3] Naval Research Laboratory, "Extendable Mobile Ad Hoc Network Emulator (EMANE)"; http://cs.itd.nrl.navy.mil/work/emane/, accessed May 23, 2017.

[4] J. Yackoski *et al.*, "RF-NEST: Radio Frequency Network Emulator Simulator Tool," *Proc. IEEE MILCOM*, Baltimore, MD, Nov. 7–10, 2011, pp. 1882–87.

[5] GNU Radio; http://gnuradio.org/redmine/projects/gnuradio/wiki, accessed May 23, 2017.

[6] REDHAWK; http://redhawksdr.github.io/Documentation, accessed May 23, 2017.

[7] LabView; http://www.ni.com/white-paper/14297/en/, accessed May 23, 2017.

[8] http://www.mathworks.com/discovery/sdr.html, accessed May 23, 2017.

[9] https://www.xilinx.com/products/silicon-devices/soc/zynq-7000.htm, accessed July 3, 2017.

[10] http://www.rtl-sdr.com/about-rtl-sdr/, accessed July 3, 2017.

[11] E. Axell *et al.*, "Spectrum Sensing for Cognitive Radio: State-of-the-Art and Recent Advances," *IEEE Signal Processing Mag.*, vol. 29, no. 3, May 2012, pp. 101–16.

[12] S. Soltani *et al.*, "Distributed Cognitive Radio Network Architecture, SDR Implementation and Emulation Testbed," *Proc. IEEE MILCOM*, Tampa, FL, Oct. 26-28, 2015.

[13] Z. Lu, Y. E. Sagduyu, and J. Li, "Queuing the Trust: Secure Backpressure Algorithm against Insider Threats in Wireless Networks," *Proc. IEEE INFOCOM*, Kowloon, Hong Kong, April 26–May 1, 2015.

[14] Y. E. Sagduyu *et al.*, "Regret Minimization-Based Robust Game Theoretic Solution for Dynamic Spectrum Access," *Proc. IEEE CCNC*, Las Vegas, NV, Jan. 9–12, 2016.

[15] Orbit testbed project, http://www.orbit-lab.org, accessed May 23, 2017.

Biographies

Yalin Sagduyu [SM] is the associate director of Network and Security at Intelligent Automation, Inc. He received his Ph.D. degree in electrical and computer engineering from the University of Maryland, College Park. His research interests include wireless networks, cognitive radio, and machine learning. He has chaired workshops at ACM MobiCom, IEEE CNS, and IEEE ICNP, served as Track Chair at IEEE PIMRC 2014, and served on the Organizing Committee of IEEE GLOBECOM 2016.

Sohraab Soltani is lead scientist at Intelligent Automation, Inc. He received his Ph.D. degree in computer science from Michigan State University. His research interests are in wireless communications, tactical networks, radio programming, cognitive radio, ad hoc and sensor networks, stochastic networking, and wireless system implementation. In this career, he has implemented a broad set of wireless (PHY, MAC, network layer) technologies on various radio platforms and executed in-lab and field tests.

Tugba Erpek is a senior research engineer at Intelligent Automation, Inc. She received her M.Sc. degree in electrical and computer engineering from George Mason University in 2007. Her research interests include wireless communications, cognitive radio, dynamic spectrum access, network protocols, and routing algorithms. She is currently pursuing her Ph.D. degree in electrical engineering at Virginia Polytechnic Institute and State University focusing on machine learning for communication systems.

Yi Shi is a senior research scientist at Intelligent Automation, Inc. His research interests include algorithm design, optimization, machine learning, communication networks, and social networks. He has been an Editor for *IEEE Communications Surveys & Tutorials*, a TPC chair for IEEE and ACM workshops, and a TPC member of many major IEEE and ACM conferences. He was a recipient of the IEEE INFOCOM 2008 Best Paper Award and the IEEE INFOCOM 2011 Best Paper Award Runner-Up.

JASON LI is a vice president and senior director at Intelligent Automation, Inc., where he has been working on research and development programs in the area of networks and cyber security. He received his Ph.D. degree in electrical and computer engineering from the University of Maryland, College Park. He has led numerous R&D programs related to protocol development for realistic and repeatable wireless networks test and evaluation, moving target defense, and cyber situational awareness.

SAVE THE DATE FOR THESE UPCOMING TRAINING COURSES!

Combining Software-Defined and Delay-Tolerant Approaches in Last-Mile Tactical Edge Networking

Iulisloi Zacarias, Luciano P. Gaspary, Andersonn Kohl, Ricardo Q. A. Fernandes, Jorgito M. Stocchero, and Edison P. de Freitas

Observing a gap in the literature about handling problems arising from the last mile TEN, and taking into account the resource-constrained devices used by troops in the field, the authors propose the joint exploration of SDN and DTN concepts to address the needs of tactical-operational networks.

ABSTRACT

Network-centric warfare is a no-way-back trend in modern military operations. The application of this concept ranges from upper-level decision making echelons to troop guidance on the battlefield, and many studies have been carried out in this area. However, most of these are concerned with either the higher-level strategic networks, that is, the networks linking the higher echelons with abundant resources, satellite communications, or even a whole network infrastructure, or high-end TEN, representing resource-rich troops in the field, with military aircraft, battleships, or ground vehicles equipped with powerful wireless communication devices and (almost) unrestricted energy resources for communication. However, these studies fail to take into account the "last-mile TEN," which comprises resource constrained communication devices carried by troopers, equipping sensor nodes deployed in the field or small unmanned aerial vehicles. In an attempt to fill this gap in the studies on battlefield networking, this article seeks to combine software-defined and delay-tolerant approaches to support the diverse range of strict requirements for applications in the last-mile TEN.

INTRODUCTION

Currently, military operations are controlled by interconnected units, from a strategic level (i.e., the higher-level centers for decision making) to the tactical-operational level involving the units on the battlefield. Battlefield networks (BNs) establish connections among these different nodes through various communication technologies, from short-range wireless to satellite links. At a tactical-operational level, the involved networks are referred to as tactical edge networks (TENs) [1], which vary depending on the capabilities of the nodes that form them. At one end, there are networks made up of high-end nodes, which are resource-rich in regard to communications technologies, high bandwidth, and (almost) no restrictions on energy consumption [2]. At the other end, networks of resource-constrained nodes, also called low-end nodes, cover the following: the last-mile BN consisting of wireless sensor nodes deployed in the field, radios carried by special force units, and small unmanned aerial vehicles (UAV) for image acquisition, among other areas [3]. These different networks support a variety of applications, ranging from elastic ones, such as file and text message transfer, to non-elastic real-time applications, such as video streaming. Applications are affected by different problems and have to meet different requirements depending on their degree of elasticity. To tackle these problems, the overall communication system that forms a part of the battlefield scenario can employ different concepts, such as software-defined networks (SDNs) and delay/disruption-tolerant networks (DTNs).

SDN is a promising paradigm, which provides flexibility to network management by separating the network infrastructure into distinct planes [4]. Each plane can be programmed to meet particular application requirements; for example, to support legacy or commercial off-the-shelf (COTS) applications without having to make intensive changes to the behavior of an application. These network adjustments may range from quality of service (QoS) configurations to the deployment of new protocols and policy enforcement in a running network. The application programming interface (API) provided by an SDN controller enables the technology to be integrated with high-level non-network systems. These systems can deploy security, media, or vendor-specific features using the SDN controller API regardless of the network protocols. The use of SDN in the BN scenario was previously explored in [2]. It aimed to improve communications between devices in a dynamic and heterogeneous military environment. The authors adopted a high-level approach to apply SDN concepts to military networks where the communication nodes are resource-rich (e.g., battleships and airplanes) using links provided by satellites, among other methods. Their main concern was to strengthen a military network at a strategic level through the flexibility provided by SDN.

DTN is a network architecture approach that aims to address the problem of lack of continuous connectivity in dynamic networks; that is, there are extended periods of link unavailability [5]. TENs often employ relay nodes (RNs) to forward data from source to destination nodes and thus acquire a larger coverage area (including rugged terrain and harsh environments), where they are usually deployed. However, RNs may be unavailable as a result of energy depletion, tech-

Digital Object Identifier: 10.1109/MCOM.2017.1700239

Iulisloi Zacarias, Luciano P. Gaspary, and Edison P. de Freitas are with the Federal University of Rio Grande do Sul; Andersonn Kohl, Ricardo Q. A. Fernandes, and Jorgito M. Stocchero are with the Brazilian Army.

nical failures, or attacks by enemy forces. In addition, due to the high mobility of some nodes (e.g., UAVs and ground vehicles), the data routes often have to be recalculated, and sometimes end-to-end connection between a given source and destination nodes does not exist, or only exists for a very short time. Long link outages lead to loss of connectivity, transmission timeouts, and routing failures. DTN tackles this problem by temporarily storing the data in RNs. These data are then forwarded to the destination when opportunistic connections show up. As outlined in [6], DTN is a technology widely explored in military networks due to their need to augment the coverage of communication networks by using aerial high-capacity backbone systems comprising manned and unmanned aircraft. Links between aircraft display more frequent intervals of outage than ground or satellite networks. However, like other approaches such as SDN usage in BNs, the DTN design described in [6] is also restricted to networks at the strategic level, which rely on resource-rich nodes, and does not cover the last-mile TEN.

Observing this gap in the literature about handling problems arising from the last-mile TEN, and taking into account the resource-constrained devices used by troops in the field, this work proposes the joint exploration of SDN and DTN concepts to address the needs of these tactical-operational networks. The proposed approach benefits from the programmability offered by SDN and the ability of DTN to handle link outages. These features are used in the context of network-centric military operations in the field, and primarily to fulfill the hard-to-meet requirements of low-end nodes. Thus, the main contribution of this article is to define an architectural solution to support the last-mile TEN by adapting and combining SDN and DTN technology.

Reviewing SDN and DTN Concepts

The SDN paradigm was initially planned as a form of technology that could be applied in wired networks. It was rapidly adopted by the research community and industry, which extended it for other types of (wireless) networks, such as satellite networks [6] and wireless sensor networks (WSNs) [7]. The paradigm proposes the separation of the data forwarding plane (or data plane) from the network control logic (control plane). The network equipment in the forwarding plane consists of simple packet forwarding devices, while the control logic of the network is implemented in a centralized entity, called the SDN controller [8]. It is noteworthy that centralized control does not imply having a physically centralized entity. Indeed, single points of failure (SPOFs) should be avoided, and redundancy in the control plane is highly desirable [4].

The clear separation between the network planes and the abstraction of the network control logic from the distributed hardware makes it easier to implement new network management directives. An external application can interact with the network through the SDN controller. Examples of external applications are load balancers, network orchestration frameworks, and business functions. The SDN architecture allows interaction between planes by defining protocol-neutral interfaces. The southbound interface allows the control plane to exchange data with the forwarding plane. The current well-known de facto standard protocol adopted by industry and used in the southbound interface is OpenFlow [7]. Communication between external applications and the control plane occurs through the northbound interface. External applications use this interface to communicate their requirements to the SDN controller or to get information about the overall state of the network. The SDN controller translates the high-level specifications from the application plane through a northbound API to low-level specifications and applies them to the data plane by means of the southbound API [4].

DTNs were originally designed to cope with problems of an interplanetary Internet (IPN). This environment is subject to problems such as large round-trip times (RTTs), limited bandwidth, errors in communication links, possible link disruptions for long intervals, and intermittent connectivity. Researchers also recommend using DTN solutions in satellite applications because satellite links share some of the difficulties found in IPN. Despite the completely different scenario, terrestrial applications also make use of DTN solutions. Examples of terrestrial applications that make use of DTN solutions are BNs (including TENs), low-density and underwater wireless sensor networks, and UAV communication systems [6].

As a means of overcoming the problems caused by link disruption and large delays in data transmission, the DTN nodes act in a store-carry-forward way [5]. Link interruptions can range from a few seconds to several hours depending on the type of application employed. The DTN nodes must temporarily store a potentially large amount of data, and the intermediary DTN nodes are responsible for it until it can be forwarded to the next DTN-capable node or the destination node [6]. The Internet Research Task Force (IRTF) proposes an architecture and protocols that address data exchange by means of DTNs. An additional layer is placed between the transport and application layers to provide the special features required by DTNs [5].

The SDN paradigm was initially planned as a form of technology that could be applied in wired networks. It was rapidly adopted by the research community and industry, which extended it for other types of (wireless) networks, such as satellite networks and wireless sensor networks.

Application Scenario

TENs are used to integrate communication devices in harsh battlefield environments. Scenarios in which TENs can be applied include, but are not limited to, borderline surveillance and the exploration of enemy-occupied areas. Access to the environment may be limited or restricted, and often there is no pre-existing communication infrastructure like cellular fourth/fifth generation (4G/5G) networks. The communication network in these scenarios is inherently very dynamic, due to the presence of mobile nodes such as small UAVs, ground vehicles, and the devices carried by soldiers [3, 9]. Small and energy-constrained devices are also combined with the TEN, such as piezoelectric sensors used to detect enemy forces [3, 10]. To benefit from the acquired data, there is a clear need for data exchange and forwarding. The importance of this data exchange and forwarding is to provide data to support commanders in their decisions, assisting in the management of the military operations [11].

The data exchanged between the TEN nodes are very diverse, ranging from simple and sin-

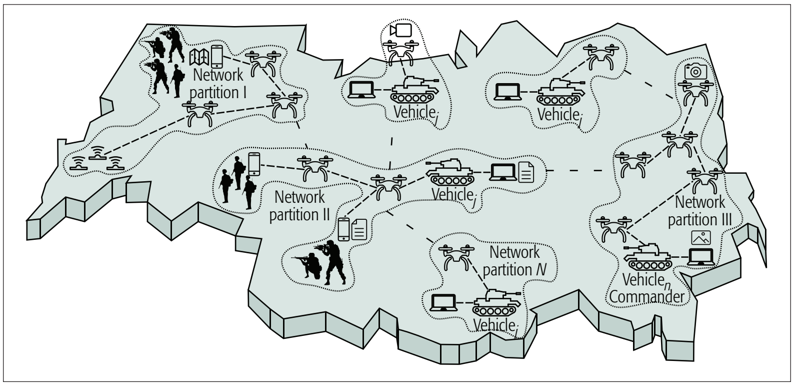


Figure 1. Last-mile TEN application scenario.

gle-valued data collected by small sensors [10] to a large amount of data generated by video capture devices (e.g., visible light and infrared cameras) and file transfers [3]. These data flows have different QoS requirements, depending on the applications [9]. For example, a file transfer application has flexible requirements related to delay, delay variation (jitter), and bandwidth. This is in contrast with a video surveillance application, which requires near-real-time transfers. A delayed video stream transmission of a surveillance application may affect the decision making and therefore hamper military operations. In this context, small UAVs can be used to fly ahead of the troops to gather information about hostile threats and enemy positioning.

Examples of the above mentioned applications in last-mile TEN are shown in Fig. 1. It is important to notice that there is not a single (big) network, but several small networks that occasionally and opportunistically interconnect with each other. For instance, a group of small UAVs launched from a ground military vehicle to video survey the terrain ahead forms its own network. This network may not be connected to the launching vehicle for a given period of time due to interference or range limits. However, this connection may eventually be restored, and a UAV closer to the ground vehicle can deliver the video that one of them has acquired. Another example, illustrated on the left side of Fig. 1, is a WSN for motion detection, which may be completely disconnected from the other nodes outside the WSN. These sensors acquire and process information about enemy movements in the area, and when a friendly node comes within range (e.g., a UAV), they supply the acquired data. Following this example, this UAV can, in turn, forward the data through a UAV-relay network to command and control (C2) applications running on mobile devices carried by troops in the field. As these applications have different requirements, they can be systematically classified so that a careful analysis can be conducted to address these specific requirements.

The dissemination of information is an essential part of C2 systems, since they directly affect the battlefield scenario [10, 11]. The use of information grids is a popular approach in network-centric warfare, because it allows situation awareness sharing by exchanging data between different applications and systems [1]. Legacy and COTS applications are also often used in C2 systems, and messages exchanged between them usually benefit from service-oriented architectures (SOAs) in TENs [12]. The SOA concept is a means of building distributed systems and a fundamental principle for federated military systems [9]. The SOA approach makes it easier to carry out data sharing between different applications and offers a flexible mechanism for reuse of already existing services [12]. TEN nodes can exchange SOA messages with each other by means of SOA-based software. For example, a C2 application handles maps running on handheld devices, which must subscribe to receive updated data from a WSN to be displayed. The use of intermediary nodes for caching the information and forwarding it over opportunistic wireless links causes delays to the communication. However, these delays do not represent a significant problem in elastic applications. For example, file transfers of photos and topographic maps among the TEN nodes belong to an elastic type of application, which can thus tolerate the store-carry-forward approach [13].

Applications that make use of a video to survey or secure an area should be able to meet stricter network requirements [14]. They are used when rapid decision making is required, which means that they are not as tolerant to network service degradation as the elastic applications. By nature, video transmission requires higher throughput, and the application cannot support a higher degree of network instability, like throughput limitations and significant latency variation (jitter). Bandwidth and latency variations result in inefficient video data transfers and low-quality video playout, with freezes and degradation of displayed images. The result is a loss of important details, which may lead to wrong decisions. The quality of images and video is crucial to the decision making process and cannot be disregarded [11]. In light of these requirements, the group of applications dealing with video transmission and continuous data flows having strict QoS network requirements can be classified as *non-elastic applications* [13].

Applications using bidirectional data transfers, such as video conferencing and voice over IP, are also valuable in networked warfare systems [14]. The network requirements imposed by these applications are very strict because the communication is an interactive process. The network should support a minimum throughput by allowing the transmission of video and/or audio and additionally meet strict requirements regarding latency variation (jitter). In these applications, end-to-end connections must exist between the points that wish to communicate, which means that temporary data storage at intermediate nodes using DTN solutions is inappropriate. Applications that have these restrictive features are classified as *interactive non-elastic applications*.

Table 1 summarizes the main characteristics of the aforementioned application classes, comparing their network requirements and suitability for making use of DTN intermediary storage.

In addition to network resource requirements, other important factors that need to be observed are the following:

- The deployment of new devices should meet strict requirements with regard to timing. Ease of configuration and fast device replacements are also important.

- In view of its high dynamicity, the network has to be prepared to promptly respond to changes in the topology, and even be able to provide self-healing properties.
- The different data flows require different QoS levels according to application classes.
- Security is a first-order requirement for military networks. Thus, the deployment of security mechanisms is essential.
- The nodes operating in TENs are often subject to energy constraints. Optimized routing algorithms and energy-aware protocols are desirable to extend the operating time of battery-powered devices.
- There are plenty of legacy applications already being deployed in military networks. The need for significant changes in these applications should be avoided. Existing applications should be included and seamlessly integrated.
- There is a need to avoid SPOFs in military systems, and for this reason link redundancy mechanisms are essential components in military networks [9].

The Proposed SDN-DTN Military Network Architecture

The last-mile TEN represents a challenging network environment that can be explained by the high mobility of the nodes, device heterogeneity, resource constraints, and specific (and hard to meet) requirements of military applications. This network connects applications with different requirements with regard to QoS, security, and reliability. When the concepts of SDN are merged with those provided by DTN, they offer the means to tackle this challenge. This section describes the objectives of the proposed architecture together with the best features offered by each approach.

By following the SDN paradigm, the forwarding devices are remotely managed by a centralized SDN controller, although this does not imply that it is a physically centralized entity. TENs must be robust and tolerate attacks so that they are able to continue operating even with damaged components. Redundancy mechanisms for the SDN controller are required to achieve the desired robustness.

The use of multiple controllers in the TEN overcomes the SPOF problem by allowing the network to continue operating when it is disrupted. By selecting a scenario in which small UAVs form a part of a TEN, each of them can take control of the network by running a controller instance. However, unnecessary control processing leads to waste of energy. On the other hand, a single instance of an active controller for the whole network will be unreachable for nodes located in different network segments. In this case, an SDN controller should be selected in each segment by running a leader election algorithm. When all the nodes return to the same segment, the leader election algorithm is run again and elects a single controller for the entire network. The elected controller might be one of the controllers previously active in one of the partitions, or even a new one running in a different node. In this case, the new controller is synchronized with the previous ones to ensure that consistency is kept with the already applied configurations. Stress should be laid on awareness of the overhead, or even the infeasibility of this election process due to the extreme dynamics of such networks. In these cases, the proposed architecture should be slightly changed and include, for instance, an out-of-band control channel. The assumption here is that the dynamics of the network is at a degree in which changes are expected to happen in minutes. This is plausible in these scenarios and ensures the feasibility of the approach.

Parameter	Application class		
	Elastic	Non-elastic	Interactive non-elastic
Network throughput usage	Low	High	Medium–high
Network latency tolerance	High	Low	Low
Jitter tolerance	High	Medium	Low
Packet loss tolerance	Low	Medium–high	Medium
DTN storage allowed	Yes	No	No
Example applications	File transfer, messaging, SOA-based applications	Video surveillance	Video and/or audio conference

Table 1. Applications classes and network requirements.

The node that runs an active instance of the SDN controller is the master in that network segment, and responsible for its control. The other nodes in the same segment are called slaves. Figure 2 shows the internal architecture of the master and slave nodes on its left and right sides, respectively. Instead of an SDN controller, the slave nodes run an SDN daemon component (on the right side in Fig. 2), which continuously surveys the current network segment to detect active SDN controllers. When an SDN controller is not found, the SDN daemon starts the leader election algorithm. When a slave node becomes a master (the new master elected node), the SDN daemon initializes the SDN controller and the DTN orchestrator components in that node (top right and top middle of the master node in Fig. 2). The other slave nodes in the TEN segment will be notified and have to update their configurations to connect to the recently elected controller.

There may be a huge number of messages exchanged between the master (SDN controller) and slave nodes, thus increasing energy consumption. To address this concern, the SDN daemon manages the migration of the SDN controller process among the nodes forming the current network segment. This management system aims to avoid battery depletion of the master node, thus increasing the autonomy of the small battery-powered devices that are often used in last-mile TENs. It should be noted that an instance of the SDN controller will run on a node when it is isolated from the remaining network. This controller will only manage this node's network interface. Also, it will watch for an opportunity to rejoin a network segment, and when this occurs, it will run the leader election algorithm and forward the temporally stored data that it may have.

The communication between the forwarding devices and the SDN controller uses the managed

The UAV on the top collects video to be sent to the commander's vehicle, located at the bottom. This task is valuable in reconnaissance missions, when a platoon arrives at a region of interest and UAVs are used to search for possible threats. The video transmission is a non-elastic application and must meet strict QoS requirements.

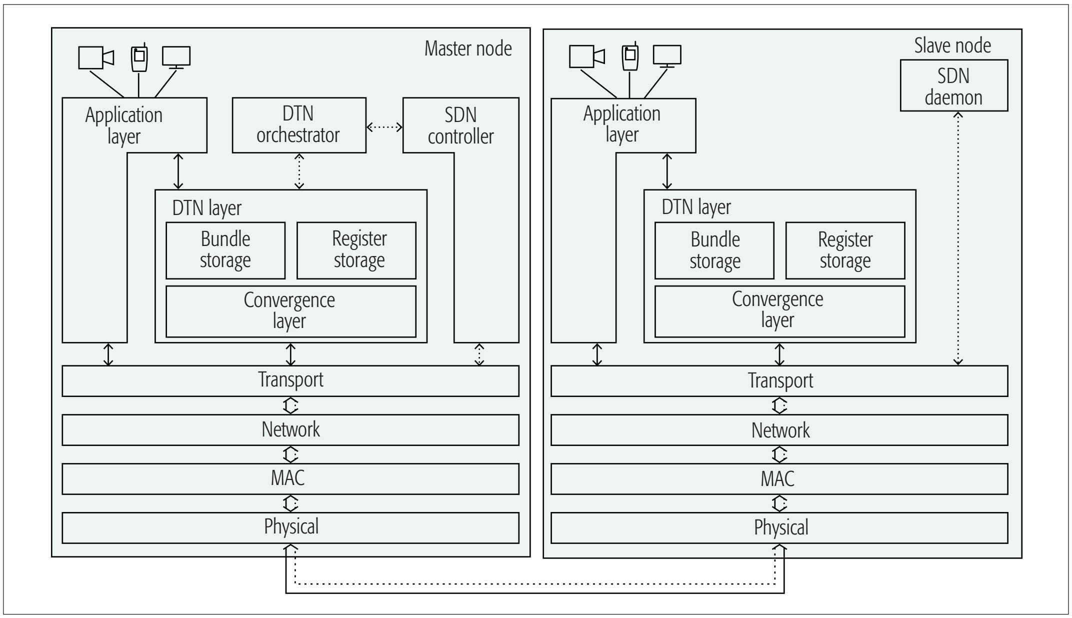


Figure 2. Combined SDN and DTN architecture.

network, and therefore an in-band controller connection [8]. The OpenFlow messages are routed to controlled devices through conventional network connections, and hence traverse the transport, network, and medium access control (MAC) layers, and finally the physical wireless interface (Fig. 2, at the bottom of both the master and slave nodes).

The DTN orchestrator coordinates the functionality of the DTN nodes. This module exchanges link information with the SDN controller in the master node, as can be observed in Fig. 2 (on the left side). The DTN orchestrator schedules data transmission between the DTN nodes on the basis of data collected by the SDN controller. The SDN controller has updated information and a "global view" of the network, that is, a global view of the network segment it is controlling. Data flows in TENs have different QoS and security requirements [2], and the SDN controller keeps track of them. The DTN orchestrator also gathers information from the DTN layer about the bundle storage state, buffer utilization, and DTN endpoints (register storage). The acquired information can be used to select a different DTN routing algorithm or forward the data so that it can be stored temporarily in another node, thus avoiding the overflow of the DTN buffer.

The DTN layer, represented in Fig. 2 (in the middle of both master and slave nodes), follows the concepts of current DTN implementations. The bundle storage component is used to store the data collected by devices and applications when the destination node is unreachable (e.g., the destination node is in another network segment or is temporarily turned off). The convergence layer links the bundle storage to the transport layer. The data stored in the bundle storage buffer are sent to the destinations or neighbor nodes using the UDP or TCP transport layer protocols. The delay-tolerant data, which are temporarily stored in nodes, are routed by means of specific DTN algorithms. There are many routing algorithms for DTNs, and a particular set of routing algorithms can be employed to route the data between the DTN nodes in TENs, such as geo-routing algorithms. The DTN orchestrator manages the routing protocol switching, and it chooses the most suitable one in accordance with internal decision algorithms and the data exchanged with the SDN controller. The flexibility provided by the SDN makes it easier to change the routing protocol while the TEN remains running.

SDN-DTN Architecture in Action

Two use cases are examined to illustrate the practical application of the proposed architecture. The first sets out a scenario that highlights some of the SDN features within network partitions supporting a non-elastic application. The second complements the first by giving a description of an alternative situation, that is, a scenario in which two different network partitions are connected, supporting an elastic application.

Use Case #1: Intra-Network Partition

The first use case handles simultaneous video streaming in an intra-network partition scenario. This situation is depicted in Fig. 1, partition III. The UAV on the top collects video to be sent to the commander's vehicle, located at the bottom. This task is valuable in reconnaissance missions, when a platoon arrives at a region of interest and UAVs are used to search for possible threats [11]. The video transmission is a non-elastic application and must meet strict QoS requirements, as shown in Table 1, or there is a risk that the user may lose significant events.

Although all the devices are in the same partition, due to the distance between the source and destination nodes, RNs must be employed. The bandwidth of an RN may be shared with many data flows from different video sources. The SDN controller has global and near-real-time information about the link usage, and thus can select appropriate paths to forward video flows that comply with the requirements. Additionally, the controller employs mechanisms to ensure fairness among concurrent video flows or evenly distribute the data streams among redundant links. For example, in an area of 4 km × 4 km in which nine UAVs and nine ground vehicles form a mesh network, from one to nine UAVs may provide video streams toward the commander's vehicle. Figure 3 shows the video playback start time (Fig. 3a), the number of video stalls (Fig. 3b) and the lengths of the stalls (Fig. 3c) for simulations performed according to the parameters provided in Table 2. The videos are compatible with military application needs [9, 14]. The video metrics were collected from the player reflecting the user experience. According to [9, 11], it is possible to infer that for TEN operating environments and the way a military operation in this level unfolds, a small number of short video stalls (around 10–15 stalls of 1–2 s each) do not harm the usage of the video. By the presented results, with more than six streams the system becomes visibly degraded. However, this degradation is due to the inherent limits of the used physical layer, as the SDN controller correctly behaved, selecting the best routes. Improvements can address this problem by changing the physical layer or using adaptive protocols to diminish the image resolution or the frame rate. Without using the proposed architecture, even with just one stream, the results are poor and, for many simultaneous streams, completely unacceptable.

Use Case #2: Inter-Network Partitions

The second use case deals with communication between heterogeneous nodes in the TEN, spread across different partitions. In Fig. 1, there is an illustration of the transmission of data acquired by WSN nodes located in partition I (on the left of Fig. 1) to a C2 application running in the vehicle in which the operational commander is following the battle unfold, in partition III (on the right in Fig. 1).

The WSN is initially isolated from the other partitions, and covers an area that is far away or difficult to access, where it collects data about enemy movements. The sensor nodes store the collected data in temporary memory or forward them to a particular sensor node that is capable of storing larger amounts of data.

A UAV performing a task nearby is assigned to collect the WSN data. When it establishes a connection with the WSN nodes, the leader election chooses a node to run the SDN controller and DTN orchestrator instances. The SDN controller updates the topology data concerning the current network partition. The flow rules applied to the forwarding devices classify the data being transmitted between the WSN nodes and the UAV. The WSN data belong to an elastic application, and thus DTN support is used. The SDN controller notifies the DTN orchestrator about the data being received and optimizes the data transfer between the DTN nodes and the UAV, by selecting links that are capable of proceeding with the data transfer. These links connect the UAV to other UAVs and/or military ground vehicles that can act as data forwarding nodes inside a given partition and between nearby partitions that are opportunistically joined.

Parameter	Used Value
Number of UAVs	9
Number of ground vehicles	9
Total simulation area	4 Km × 4 Km
UAV moving area	4 Km × 1 Km
UAV communication radius (R_u)	360m
Ground communication radius (R_v)	650m
Video size	960 × 540 pixels
Video frame rate	30 f/s
Video codec	H.264
Video length	60 seconds
Number of runs per number of streams being served	33 runs
Wireless standard	802.11g
Frequency range	2.4 – 2.485 GHz
Data rate	Up to 54Mb/s
Simulation environment	Mininet-WiFi
Mobility model	Random Waypoint

Table 2. Parameters used in the simulations.

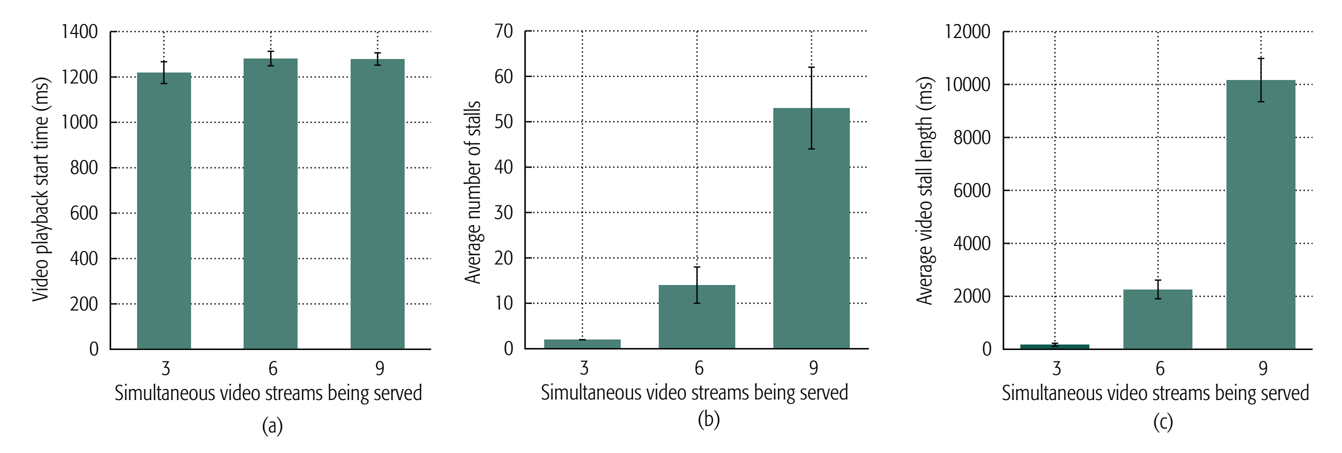


Figure 3. Simulations employing simultaneous video transmissions: a) playback start time; b) average number of stalls per video stream; c) average video stall length per video stream.

Observing advances in both SDN and DTN paradigm, a promising approach to combine the best of these to address last-mile TEN is proposed in this work. The architecture shows features capable of tackling this difficult operational scenario. It should also be stressed that other domains may benefit from this architecture.

As the UAV continues its movement around the area, a connection is eventually made with another partition. Again, the leader election algorithm selects a node to host the SDN controller and DTN orchestrator instances. The DTN orchestrator is aware of the data acquired from the WSN and interacts with the SDN controller to choose the best route to forward the data stored by the DTN layer toward the destination node. The SDN controller installs flow rules in the forwarding devices to allow the data to be transmitted via partition II until it reaches the destination in partition III. Preliminary results from the experiments within this scenario show that the proposed approach is able to successfully deliver 100 percent of the data sent by the WSN. On the other hand, the same network without the proposed SDN-DTN approach has results with a success rate for delivered data ranging from 44 to 55 percent.

Another possible way of using the WSN data in this scenario would be similar to mobile micro-clouds [15], in which the data could be processed by services placed on nodes inside partition I, after the UAV that gathered the data from the WSN rejoins this partition.

Conclusion

The BN is a challenging communication environment in which very different concerns have to be considered. Attempts to address these concerns have involved exploring approaches based on SDN or DTN. Despite the features that these two paradigms provide, they cannot address all these concerns on an individual basis. Moreover, few proposals have examined them in the last-mile TEN, which is the most challenging part of the BN due to its high mobility and constrained resources.

Observing advances in both the SDN and DTN paradigms, a promising approach to combine the best of these to address last-mile TEN is proposed in this work. The architecture shows features capable of tackling this difficult operational scenario. It should also be stressed that other domains may benefit from this architecture, such as networks designed to assist disaster relief operations, which face similar harsh operational conditions.

Prospective Directions

Despite the contribution presented in this work, many challenges remain to be solved by academia and industry. Solutions to keep network topology consistency when facing highly dynamic situations are under development, and mechanisms to migrate the SDN controller between nodes connected by restricted bandwidth wireless links require further work. Network signaling in low-end nodes needs improvements to efficiently apply the SDN paradigm regarding the concern about the energy consumption of these nodes. The development of adaptive protocols to change the video rate and resolution according to network conditions is an interesting topic that also requires effort to evolve. Also, the integration of the proposed architecture with existing military network solutions needs further research, mainly related to network security.

The development of a suite of protocols specifically designed for the proposed architecture and envisaged applications might be considered as a promising future work direction. Also, improvement of the leader election mechanisms to cope with the network resource constraints and the recurrent changes in network topology is still a relevant topic to be explored. Further, the adaptation of this approach to different contexts, such as emergency networks, and the exploration of edge computing and micro-clouds to complement and extend the proposal also deserve examination.

Acknowledgments

The authors thank Marcelo A. Marotta for the valuable discussions, Ramon. R. Fontes for developing the Mininet-Wifi emulator and the support for its usage, and the State of Rio Grande do Sul Research Foundation (FAPERGS Proj. no. 2240-2551/14-2SIAFEM) for the financial support to perform this work.

References

[1] M. Tortonesi *et al.*, "Enabling the Deployment of COTS Applications in Tactical Edge Networks," *IEEE Commun. Mag.*, vol. 51, no. 10, Oct. 2013, pp. 66–73.

[2] J. Nobre *et al.*, "Toward Software-Defined Battlefield Networking," *IEEE Commun. Mag.*, vol. 54, no. 10, Oct. 2016, pp. 152–57.

[3] D. Orfanus, E. P. De Freitas, and F. Eliassen, "Self-Organization as a Supporting Paradigm for Military UAV Relay Networks," *IEEE Commun. Lett.*, vol. 20, no. 4, Apr. 2016, pp. 804–07.

[4] J. Wickboldt *et al.*, "Software-Defined Networking: Management Requirements and Challenges," *IEEE Commun. Mag.*, vol. 53, no. 1, Jan. 2015, pp. 278–85.

[5] K. Fall and S. Farrell, "DTN: An Architectural Retrospective," *IEEE JSAC*, vol. 26, no. 5, June 2008, pp. 828–36.

[6] R. Amin *et al.*, "Design Considerations in Applying Disruption Tolerant Networking to Tactical Edge Networks," *IEEE Commun. Mag.*, vol. 53, no. 10, Oct. 2015, pp. 32–38.

[7] T. Luo, H. P. Tan, and T. Q. S. Quek, "Sensor OpenFlow: Enabling Software-Defined Wireless Sensor Networks," *IEEE Commun. Lett.*, vol. 16, no. 11, Nov. 2012, pp. 1896–99.

[8] H. Kim and N. Feamster, "Improving Network Management with Software Defined Networking," *IEEE Commun. Mag.*, vol. 51, no. 2, Feb. 2013, pp. 114–19.

[9] P. Bartolomasi *et al.*, "NATO Network Enabled Capability Feasibility Study," NATO Consultation, Brussels, Belgium, tech. rep. 2.0, Oct. 2005.

[10] A. Kott, A. Swami, and B. J. West, "The Internet of Battle Things," *Computer*, vol. 49, no. 12, Dec. 2016, pp. 70–75.

[11] R. Fernandes, M. R. Hieb, and P. Costa, "Levels of Autonomy: Command and Control of Hybrid Forces," *Proc. 21th C2 in a Complex Connected Battlespace*, Sept. 2016, pp. 1–16.

[12] M. R. Brannsten *et al.*, "Toward Federated Mission Networking in the Tactical Domain," *IEEE Commun. Mag.*, vol. 53, no. 10, Oct. 2015, pp. 52–58.

[13] J. F. Kurose and K. W. Ross, *Computer Networking: A Top-Down Approach*, 6th ed., ser. Always Learning, Pearson, 2013.

[14] J. Nightingale *et al.*, "Reliable Full Motion Video Services in Disadvantaged Tactical Radio Networks," *Proc. 2016 Int'l. Conf. Military Commun. and Info. Systems*, May 2016, pp. 1–9.

[15] S. Wang *et al.*, "Emulation-Based Study of Dynamic Service Placement in Mobile Micro-Clouds," *Proc. 2015 IEEE MILCOM*, Oct. 2015, pp. 1046–51.

Biographies

Iulisloi Zacarias (izacarias@inf.ufrgs.br) is an M.Sc. student in computer networks at the Federal University of Rio Grande do Sul (UFRGS), Brazil. He achieved his Bachelor's degree in Information Systems at the Federal University of Santa Maria (UFSM), Brazil, 2015. Currently, his research interests are related to wireless networks, drone networks, software-defined networks, ad hoc networks, and the Internet of Things.

Luciano Paschoal Gaspary (paschoal@inf.ufrgs.br) holds a Ph.D. in computer science (UFRGS, 2002). He is deputy dean and associate professor at the Institute of Informatics, UFRGS. He has been involved in various research areas, mainly computer networks, network management and computer system

security. He is an author of more than 120 full papers published in leading peer-reviewed publications. In 2016, he has been appointed as a Publications Committee member of the IEEE SDN initiative.

Andersonn Kohl (kohl@cds.eb.mil.br) has a postgraduate degree in military sciences from EsAO (1998), a communication engineering degree from IME (1996), and a Bachelor's in military sciences from AMAN (1990). He is a Colonel Military Engineer in the Brazilian Army with much experience in the development and deployment of communication and software defense systems, and is currently the director and senior researcher of the C2 Division at CDS in transition to a new position as assistant to the Army vice-chief of Information Technology and Communications.

Ricardo Queiroz de Araujo Fernandes (ricardo@cds.eb.mil.br) attended a postdoctoral program in command and control at GMU (2016) and has a D.Sc. degree in Informatics from PUC-Rio, Brazil (2012), an M.Sc. in systems and computing from IME (2009), an M.Sc. in pure mathematics from UFRGS (2007), a postgraduate degree in military sciences from EsAO (2009), and a systems and computing engineering degree from IME (2002). Currently he is responsible for the research on command and control at CDS.

Jorgito Matiuzzi Stocchero (stocchero.jorgito@eb.mil.br) has an M.Sc. degree in electrical engineering by COPPE/UFRJ (2004), an M.B.A. in politics and strategy by FGV/RJ (2016), a postgraduate degree in military sciences from ECEME (2016 and 2012), a communication engineering degree from IME (1996) and a Bachelor's in military sciences from AMAN (1990). He has worked on important projects during his career in the Brazilian Army, such as SIVAM and SisTEx. Currently, he is a Colonel assigned as vice-commander of the Telematics Integrated Center.

Edison Pignaton de Freitas (epfreitas@inf.ufrgs.br) has a Ph.D. in computer science and engineering from Halmstad University, Sweden (2011), an M.Sc. in computer science from UFRGS (2007), and a computer engineering degree from the Military Institute of Engineering (2003). Currently he holds an associate professor position at UFRGS, acting in the Graduate Programs on Computer Science (PPGC) and Electrical Engineering (PPGEE) developing research mainly in the following areas: computer networks, real-time systems, and unmanned aerial vehicles.

MILITARY COMMUNICATIONS

Inband Full-Duplex Radio Transceivers: A Paradigm Shift in Tactical Communications and Electronic Warfare?

Taneli Riihonen, Dani Korpi, Olli Rantula, Heikki Rantanen, Tapio Saarelainen, and Mikko Valkama

Inband FD operation has great potential in civilian/commercial wireless communications, because it can as much as double transmission links' spectral efficiency by exploiting the new-found capability for STAR that is facilitated by advanced SIC techniques. The authors survey the prospects of exploiting the emerging FD radio technology in military communication applications as well.

Abstract

Inband FD operation has great potential in civilian/commercial wireless communications, because it can as much as double transmission links' spectral efficiency by exploiting the new-found capability for STAR that is facilitated by advanced SIC techniques. This article surveys the prospects of exploiting the emerging FD radio technology in military communication applications as well. In addition to enabling high-rate two-way tactical communications, the STAR capability could give a major technical advantage to armed forces by allowing their radio transceivers to conduct electronic warfare at the same time that they are also receiving or transmitting other signals at the same frequency bands. After comprehensively introducing FD transceiver architectures and SIC requirements in military communications, this article outlines and analyzes all the most promising defensive and offensive applications of the STAR capability. It is not out of the question that this disruptive technology could even bring about a paradigm shift in operations at the cyber-electromagnetic battleground. At least, forward-looking innovators in the military communications community would right now have a window of opportunity to engage in original, potentially high-impact scientific research on FD military radio systems, which we would like to spur on by this speculative tutorial article.

Introduction

Inband full-duplex (FD) wireless communication [1, 2] means that a radio device is receiving and transmitting information signals at the same time and at the same center frequency, as opposed to conventional half-duplex (HD) operation. Especially, to avoid misconception, one should note that neither time-division duplexing nor frequency-division duplexing (TDD nor FDD) is regarded as "real" FD operation when adopting the contemporary terminology despite them allowing simultaneous two-way conversation, because the perspective is shifted to spectrum usage at the physical layer. While the prospects of FD radio technology in civilian/commercial applications are already largely understood, this article proceeds to its novel military applications. To that end, one should again note that all off-the-shelf "full-duplex military radios" employ TDD or FDD (or both), so they are not actually FD radios in the scope of this article.

In general, when extending beyond the plain communication context, prospective military FD radios will have the progressive capability for simultaneous transmission and reception (STAR) by which they can conduct electronic warfare at the same time that they are using the same frequency band for communication or perform an electronic attack with simultaneous signals intelligence, as shown in Fig. 1. It is quite obvious that, by utilizing this superpower, armed forces could gain a major technical advantage over an opponent that does not possess similar technology. However, viable FD operation relies on efficient antenna isolation and self-interference (SI) cancellation [3, 4], because the strong electromagnetic field radiated by an FD transceiver leaks back to its own receiver circuitry, interfering with the reception of remote signals of interest that are, after wireless propagation, usually much weaker than the local transmission.

The origins of FD communications date back to circa 2008–2010 when the research idea popped up simultaneously and independently in various institutes around the world [3, 5, 6, references therein]. Currently, the research field has gained solid stature and is still receiving increasing attention, while, after many credible academic prototypes, the first commercial products are now under development. In contrast, scientific discourse on the military applications of the FD radio technology is in its absolute infancy in the open literature. To the best of our knowledge, so far the only explicit and elaborate references to FD military radios are three sentences on jammers in [3] and the initial conference presentation of our vision [7]. By this article, we aim to bring forward the prospects of FD military radios in order to induce more interest in this emerging research topic in the scientific community.

The open literature has discussed two specific FD concepts that implicitly relate to military systems through some radars [1] and so-called physical layer security [8]. In particular, continuous-wave (CW) radars are inherently based on the STAR capability [1], and researchers in the field of information theory have recently begun to develop the Shannon theory of communication links, where FD receivers

Digital Object Identifier:
10.1109/MCOM.2017.1700220

Taneli Riihonen and Olli Rantula are with Aalto University School of Electrical Engineering; Dani Korpi and Mikko Valkama are with Tampere University of Technology; Heikki Rantanen is with the Finnish Defence Research Agency; Major Tapio Saarelainen is with the Research Centre of the Finnish Army Academy.

0163-6804/17/$25.00 © 2017 IEEE

hinder eavesdropping by simultaneously broadcasting jamming signals [8]; studies in the latter context almost never explicitly mention the potential military use. However, the conceivable applications of the STAR capability in tactical radios and electronic warfare are actually much more ample and diverse than the earlier research ever realized.

In what follows, we first present an overview of general FD transceiver architectures developed originally for civilian/commercial wireless communications and extrapolate their requirements for military radios. While state-of-the-art FD radios can achieve up to 100 dB of self-interference cancellation (SIC), their effective use in military systems likely requires even more; moreover, usage in a battlefield sets special requirements for extreme robustness to electronic warfare, not to mention that they may operate at high or very high frequency (HF or VHF) instead of commercial cellular mobile radio bands. Thus, practical military scenarios are rather different from academic laboratories, where the FD technology is already demonstrated to be feasible for nonmilitary use at upper ultra high frequency (UHF) bands. Nevertheless, we believe that the same transceiver architectures and advanced SIC techniques at large can still be used for successful military STAR operation once they have been re-engineered carefully.

We then analyze the potential defensive and offensive applications of FD military radios, including those shown in Fig. 1. The STAR capability is used for defense in the form of a "radio shield" that protects its operator from an opponent. In fact, the jamming scenario postulated in [3] is a specific example of protective applications, but we can discover many others as well. In the offensive applications, the radio operator uses the STAR capability for attacking an opponent. For example, it is reasonable to envision that an attacker could send jamming to force an opponent to increase its transmission power and thus facilitate its own simultaneous signals intelligence, for example, locating the used frequency band and transmitters or intercepting communication.

In summary, we could be witnessing a paradigm shift in tactical communications and electronic warfare provided that the military communications community solves the following two research problems related to this potentially disruptive technology:

- How do we implement the STAR capability for military applications in the first place?
- What are the most suitable ways to exploit FD radios in cyber-electromagnetic battles?

Eventually, FD radios may even become de rigueur for modern troops whenever an opposing side possesses corresponding technology, necessitating rethinking of communication procedures and tactics as a countermeasure. While our study is only one of the very first steps toward addressing the above questions, we expect that this emerging open research area affords ample opportunities for making original, potentially high-impact contributions in the coming years.

Origins in Nonmilitary Communications

In the nonmilitary context, the FD operation has mainly been considered for inband relaying [5], or for boosting the capacity of the existing communication systems [3, 6]. In both of these cases, suppressing the SI is the main research challenge, and consequently a large body of literature has been produced regarding the different SIC techniques (see, e.g., [1, 2, references therein]). In the context of inband relaying, the used SIC methods might somewhat differ from those used in the more generic bidirectional data transfer applications. In particular, different spatial suppression schemes have been widely considered for multiple-input multiple-output relays [5], while RF and/or digital domain cancellation has been the prevalent choice for the generic FD devices [6]. Since such cancellation schemes can be readily applied to all types of transceivers, including relays and radars, they are the focus of this section.

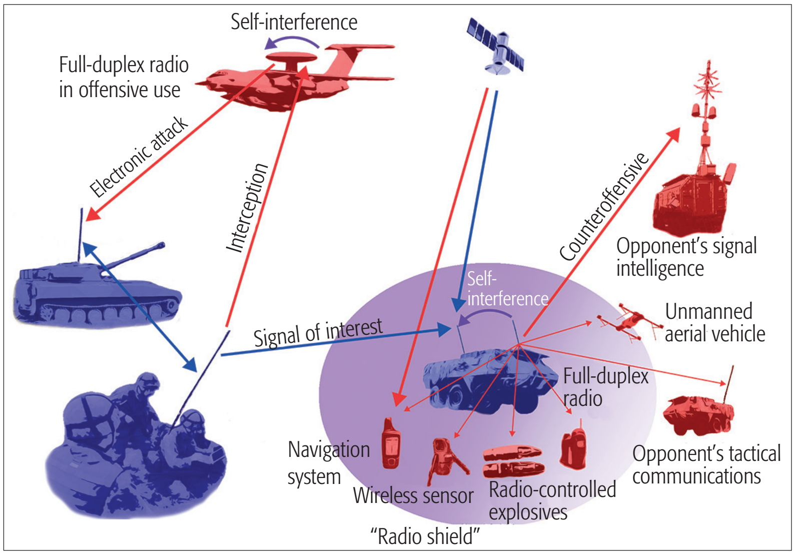


Figure 1. Conceptual view of using inband full-duplex radio transceivers in tactical communications and electronic warfare.

Self-Interference Cancellation in Full-Duplex Radios

A generic illustration of an inband FD device is shown in Fig. 2. It includes the various alternative SIC solutions, which differentiate the considered FD transceiver from a legacy HD transceiver. In principle, SIC is simply done by subtracting the device's own transmit signal from the received signal, although after appropriate modifications to ensure that the cancellation signal resembles the true SI as closely as possible. Figure 2 also demonstrates that in FD devices the transmitter and the receiver are never truly independent due to the SI and the corresponding SI cancellers. Through them, the transmit (TX) and receive (RX) chains are essentially connected from the signal processing perspective, a fundamental paradigm shift from the HD radios where the TX and RX parts are more or less independent.

Considering then the SI suppression mechanisms in Fig. 2, the antenna interface provides the first stage of isolation between the transmitter and the receiver. There are two widely considered alternatives for providing the passive isolation in this interface: using a shared TX/RX antenna together with a so-called *circulator* or simply using different TX and RX antennas. In the former case, the circulator provides the necessary passive SI isolation, attenuating the SI by roughly 20 dB, as reported in [9] for the 2.4 GHz industrial, scientif-

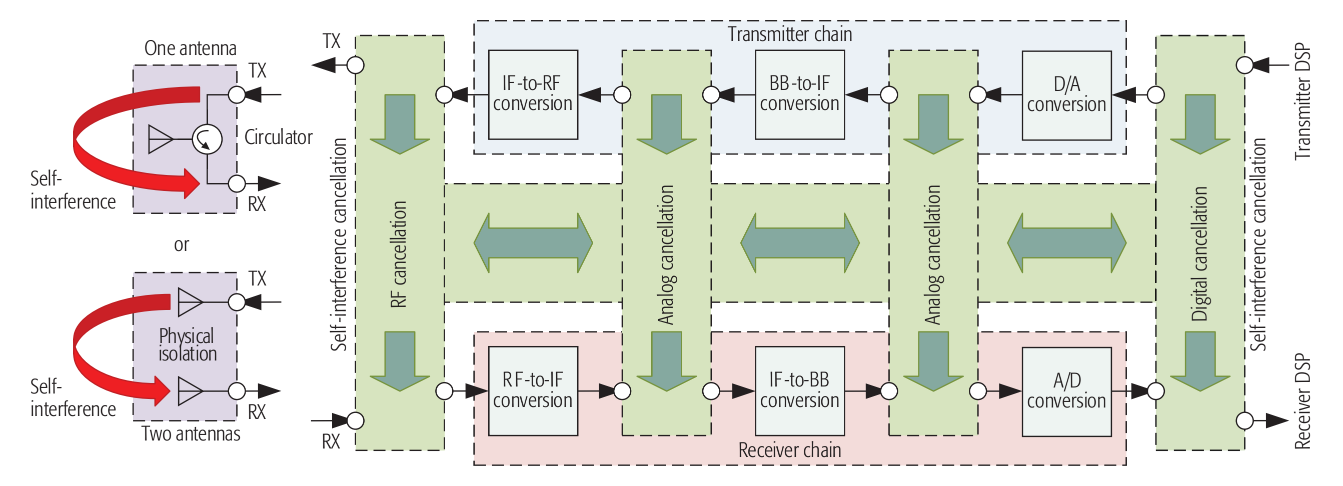


Figure 2. Generic illustration of an inband full-duplex transceiver with various self-interference cancellation solutions.

ic, and medical (ISM) band. Alternatively, if separate TX and RX antennas are employed, physical isolation is provided simply by the propagation path loss, with roughly 40 dB of isolation typically reported for the ISM band [6].

After the passive suppression, further *active cancellation* is still required before the RX chain to protect the delicate receiver circuitry. There are two prevailing techniques for such RF cancellation: one where the transmitter output signal is used to generate the RF cancellation signal [9], and one where a separate transmitter is used to upconvert a digitally generated RF cancellation signal [6]. The benefit of the former option is that there is no need to explicitly model the transmitter-induced impairments as they are already included in the RF cancellation signal, while the auxiliary transmitter based procedure profits from the fact that most of the processing can be carried out in baseband (BB) on the digital domain. The downside of using an auxiliary transmitter is the various imperfections in the main transmitter, which remain unaffected by the RF cancellation. Altogether, the RF canceller can be expected to provide 40–50 dB of SI suppression, depending on the bandwidth [9].

If employing a superheterodyne architecture in the transceiver, further analog cancellation can also be performed in the intermediate frequency (IF). In addition, also analog cancellation in the BB could be considered to decrease the SI level before the analog-to-digital (A/D) conversion. Reducing the power of the SI as much as possible before the A/D interface is highly beneficial as then a smaller number of bits for the A/D converter is sufficient to still accurately reconstruct the signal of interest in the digital domain. However, it should be noted that typically the RF canceller alone can suppress the SI sufficiently for the A/D conversion [9].

Finally, the SI remaining after the RF/analog cancellation stages is then suppressed in the digital domain by a *digital canceller*. In essence, a typical digital canceller regenerates the residual SI based on the *original transmit data*, using some predefined signal model. This means that the task of the digital SI canceller is in practice to estimate the unknown parameters of the signal model, reconstruct the observed SI using the estimated parameters, and subtract the obtained cancellation signal from the received signal. By utilizing advanced nonlinear signal models, the digital canceller can attenuate the SI by as much as 25 dB [9], thereby cancelling it almost perfectly.

Altogether, a total SI suppression of roughly 90 dB is reported in [9] for a bandwidth of 80 MHz, where a circulator, an RF canceller, and a nonlinear digital canceller are used to cancel the SI. In [4], on the other hand, 100 dB of SI cancellation is obtained over 20 MHz by utilizing active cancellers and a high-isolation relay antenna. Furthermore, and perhaps more importantly, the architectures in [4, 9], together with various other demonstrator implementations are capable of suppressing the SI close to the level of the receiver noise floor, and hence doubling the spectral efficiency is attainable. This is a promising result since it is reasonable to presume that similar transceiver designs can be used for military applications as well, albeit with somewhat different and more demanding requirements. For instance, the results on SI cancellation reported thus far are typically obtained at upper UHF bands, while the much lower frequencies at HF, VHF, and lower UHF bands are widely used by military systems instead. Although there are practically no existing works investigating STAR operation on these bands, it is likely that the same SIC solutions can be successfully applied there. However, the amount of physical isolation between the TX and RX chains is likely to be somewhat less with these lower frequencies, increasing the performance requirements for the active SIC stages.

STAR for Doubled Spectral Efficiency

The main reason for FD operation in civilian/commercial systems is the increased spectral efficiency, which results in higher data rates without increasing the bandwidth of the system. This is highly desirable nowadays due to the heavy congestion of the available spectrum, a challenge that military communication systems are also facing. In the simplest case of two FD-capable nodes engaging in bidirectional data transfer, FD operation doubles the spectral efficiency [1, 2]. However, such a symmetric point-to-point link is not a very practical scenario, and in actual real-world applications the FD devices must also provide an improvement in spectral efficiency under much more diverse conditions. For this reason, various deployment scenarios have been suggested to utilize the FD capability in different ways.

Perhaps the most intuitive application for an FD device is to use it as an inband relay or a gap filler [4, 5]. In this case, the FD transceiver merely retransmits the signal it receives, meaning that the communication scenario is fully symmetric as the same amount of time is used for both transmission and reception. Such symmetry is well suited for FD devices since then their STAR capability is utilized to the fullest extent.

Another widely considered scenario is an FD-capable base station serving half-duplex mobile users [1, 2]. Such a base station could serve uplink and downlink mobiles at the same time using the same frequency band, greatly increasing the spectral efficiency of the network. However, this type of a deployment scenario is already a bit more problematic since the uplink transmissions will in fact interfere with the downlink signals, thereby decreasing the downlink data rate. Consequently, a twofold improvement in the spectral efficiency might not be obtainable under all circumstances. Thus, further research is still needed for determining the feasibility and benefits of deploying FD nodes on a network level, in both nonmilitary and military contexts.

Continuous-Wave Radar

A very specific field, where the STAR capability has already been used since at least the 1940s, is CW radars (as opposed to pulsed radars) [1]. In this case, the direct leakage between a device's own transmitter and receiver must be efficiently suppressed, while echoes from targets must be successfully received, meaning that some form of SI cancellation is needed [1]. In fact, a CW radar using one or two co-located antennas (monostatic or pseudo-bistatic) is technically similar to a one- or two-antenna FD radio. In the former, circulators can be used to suppress the direct leakage from the transmitter, while separate antennas are used to provide the necessary isolation in the latter.

Acknowledging that radar systems typically require less isolation than FD data transfer applications, it is clear that the state-of-the-art SIC solutions, achieving 90–100 dB of SI suppression, could readily be used for low-power military radar applications. Furthermore, even though these radars typically use much higher frequencies than the reported FD prototypes, many of the SI cancellation solutions could potentially be applied also to millimeter-wave systems. Consequently, the current FD prototypes already provide many features necessary for military CW radars.

Full-Duplex Military Radio Transceivers

Requirements for Military Radios

When considering radios intended for military use, the requirements for the hardware and system-level details, listed in Table 1, are different compared to commercial applications. Military applications face many challenges, including the requirements for bandwidth, latency, stability, security, connectivity, and especially reliability [10–12]. Furthermore, impeding the radio communication of the enemy forces is also an important aspect that should be given some attention [13].

Requirements	Possibilities
Time-variant topology, stability, latency, connectivity	Communications middleware for constant radio topology awareness
Bandwidth	Full-duplex, cooperation between the radar and communication systems
Tolerance for jamming, secured communications	CSS, DSSS, FHSS, TH
Jamming and interception capability	Simultaneous full-duplex communications and jamming/interception

Table 1. Requirements and possibilities of full-duplex radios in military systems.

Perhaps the most distinguishing feature of the wireless systems designed for military use is their distributed and dynamic nature [12, 14]. This means that the network topology is heavily time-variant, and the different radios must be capable of constantly updating their knowledge regarding their nearby peers. Such a stringent requirement on topology awareness calls for some sort of a communications middleware approach where each radio should be capable of listening for the relevant information, while also informing friendly radios of its presence.

Military radios must also be tolerant to jamming or spoofing attacks, where a strong interfering signal is maliciously transmitted to disturb the data communication [3, 11, 14, 15]. Namely, since constant situational awareness is an essential requirement in the modern military context, each transceiver must be capable of delivering and receiving at least some data, even when there is a strong interfering signal present. Furthermore, it would be greatly beneficial if a transceiver was capable of simultaneously communicating and jamming enemy nodes on the same frequency, a feature that could be facilitated in a relatively straightforward fashion by the STAR capability as envisioned in this article. For instance, a remarkable battlefield application could be spoofing or jamming opponents' satellite navigation receivers without affecting one's own positioning.

In addition, a high security level within the network is required in military applications, meaning that the transmitted data must be encrypted by some means [14]. A variety of approaches exist for achieving this, such as chirp spread spectrum (CSS), direct-sequence spread spectrum (DSSS), frequency-hopping spread spectrum (FHSS), and time-hopping (TH). The tactical data link (TDL) network standard Link 16 has become the major information channel within the military communication systems of the U.S. Joint Services and forces of NATO [14]. Link 16 utilizes FHSS for improving immunity to jamming and introducing redundancy, although it is based on legacy HD transceivers.

Further limitations are caused by congestion of the available spectrum, which means that the spectral efficiency of the military networks must be as high as possible so that all the communication needs can be fulfilled without compromising reliability and security requirements [12]. In legacy systems, this has been achieved by high spectral reuse, efficient waveforms, and prioritizing the information that is disseminated within the network. In the future, the spectral efficiency can be further improved, for instance, by improving the

In the future, the spectral efficiency can be further improved, for instance, by improving the co-operation between the radar and communication systems, or by utilizing some of the recent advances in transceiver design, such as inband FD communications.

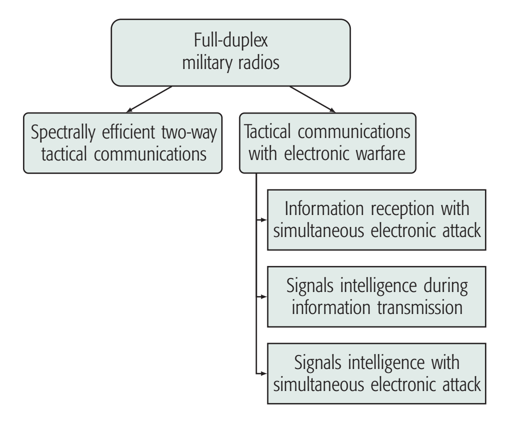

Full-duplex military radios
Spectrally efficient two-way tactical communications
Tactical communications with electronic warfare
Information reception with simultaneous electronic attack
Signals intelligence during information transmission
Signals intelligence with simultaneous electronic attack

Figure 3. Classification of applications for inband FD radios in military communications.

cooperation between the radar and the communication systems, or by utilizing some of the recent advances in transceiver design, such as inband FD communications [9].

Using Full-Duplex Radios in Military Networks

When envisioning the usage of inband FD transceivers in military communication networks, the different requirements above must be carefully considered. Table 1 shows how FD transceivers can help in dealing with these requirements, while some of the potential application modes for inband FD radios are illustrated in Fig. 3. First, inband FD communications helps in coping with the scarcity of the available bandwidth, since it can potentially provide a two-fold increase in spectral efficiency [1, 9]. This is obviously a crucial advantage in helping to ensure the situation awareness and the tactical communication capabilities under all circumstances. In this regard, many prototype implementations can already cancel the SI by a sufficient amount to realize the throughput improvements [4, 9].

In terms of the distributed nature of the network, the FD capability also allows for more efficient searching of nearby radios, since it facilitates simultaneous transmission and sensing. This has already been investigated in the context of cognitive radio systems, and shown to be feasible. Furthermore, the capability to cancel a device's own transmission also means that a jamming signal can be emitted while receiving useful data [3]. Therefore, it is clear that the FD capability creates some new possibilities beyond the improvements in spectral efficiency.

Military Communication Applications for Full-Duplex Radios

Let us consider a cyber-electromagnetic battle, where two opposing teams (blue and red) operate on the same frequency band for tactical communications and/or electronic warfare. The band can be used for transfer of information (e.g., voice, data, or an activation signal) over a link between two radios in either team and signals intelligence or an electronic attack that targets a radio in the other team. Plain two-way FD information transmission without electronic warfare is not considered herein, because it is already widely studied in the nonmilitary context, although the technology could be advantageous for facilitating high-rate tactical communications as such.

Full-Duplex Radios Only in the Blue Team

We first assume that only blue radios can operate in the FD mode, and the red team does not possess such technology. As shown in Figs. 4a–4e, we can identify five different battlefield scenarios when both teams have one or two radios, and they can be used for receiving either a communication signal or an interception signal and transmitting either a communication signal or a jamming signal.

Jamming against Communication: In the application of Fig. 4a, both teams use the same frequency band for their communications. In a conventional case without any FD radios, the blue and red teams' communication links would achieve signal-to-noise ratios (SNRs) of SNR_{bb} and SNR_{rr}, respectively. The STAR capability allows the blue receiver to transmit a jamming signal, causing extra interference to the red receiver at the cost of suffering from residual SI. Thus, the blue and red teams' communication links achieve signal-to-interference-plus-noise ratios (SINRs) of SINR_{bb} and SINR_{rr}, respectively, for which obviously $\text{SINR}_{\text{bb}} < \text{SNR}_{\text{bb}}$ and $\text{SINR}_{\text{rr}} < \text{SNR}_{\text{rr}}$ due to the fact that jamming is harmful for both teams. However, in principle, the known SI signal can be suppressed more efficiently than the unknown jamming signal so that $\text{SINR}_{\text{bb}}/\text{SNR}_{\text{bb}} \gg \text{SINR}_{\text{rr}}/\text{SNR}_{\text{rr}}$. Hence, it actually may be worthwhile for the blue team to tolerate some self-inflicted performance loss in order to have a much bigger impact on the red team's communications.

Jamming against Interception: The application of Fig. 4b is similar to the one above except that the jamming signal is now used as a countermeasure for interception. With jamming, the SINR for intercepting the blue communication link in the red receiver is given by SINR_{br} when the corresponding SNR without jamming is SNR_{br}. The information rate of the blue communication link is the same as in the above scenario, while part of the information leaks from the blue transmitter to the red receiver. Obviously, if $\text{SNR}_{\text{br}} > \text{SNR}_{\text{bb}}$ (e.g., the red receiver is closer to the blue transmitter than the blue receiver), fully covert transmission is impossible with conventional HD technology. In contrast, it is possible to achieve $\text{SINR}_{\text{br}} < \text{SINR}_{\text{bb}}$ with STAR operation even if $\text{SNR}_{\text{br}} > \text{SNR}_{\text{bb}}$. Thus, the blue link gains electromagnetic camouflage despite its total transmission rate being decreased.

Simultaneous Interception and Communication: In the application of Fig. 4c, the blue communication link uses the STAR capability for simultaneous interception. The SNR for interception would be SNR_{rb} without simultaneous information transmission, while it decreases to SINR_{rb} in FD operation due to residual self-interference. It should be especially noted that performing simultaneous interception with information transmission does not affect the blue team's own rate, so it comes at no cost during operation if the transceiver has the STAR capability. Thus, it is always worthwhile to do as long as SINR_{rb} is reasonably large such that the chances for interception are non-negligible in the first place.

Simultaneous Interception and Jamming: The application of Fig. 4d employs two FD radios for simultaneous interception and jamming in addition to communication, while the corresponding case with only one FD radio is shown in Fig. 4e. The blue team transmits jamming to the red team's receiver in order to decrease its link quality from SNR_{rr} to SINR_{rr}, which also decreases the link quality for interception from SNR_{rb} to SINR_{rb}. However, the red team may try to compensate the jamming by increasing transmission power to achieve link quality $\mathrm{SINR}'_{rr}(\mathrm{SINR}'_{rr} > \mathrm{SINR}_{rr})$ by which the link quality for interception increases to SINR'_{rb}. It is possible that $\mathrm{SINR}'_{rb} > \mathrm{SNR}_{rb}$, that is, it may be worthwhile to tolerate some self-interference in order to gain back much more from the opponent's countermove.

Full-Duplex Radios in Both Teams: There are many more battlefield scenarios when the red team also employs FD radios, as shown in Fig. 4f. We see that in the above cases the red team suffers from a technical disadvantage, because they do not possess the FD technology. For example, when the blue team is performing jamming, the smart countermove from the red team would be to launch jamming against potential interception if increasing transmit power is necessary.

Numerical Results for Jamming against Communication and Interception

Let us continue with the battlefield scenarios discussed above and illustrated in Figs. 4a and 4b. In particular, we simulate the performance of the red receiver when it is receiving a communication signal from the red transmitter or trying to intercept the blue transmission, respectively. The study assumes operation at the 2.4 GHz ISM band instead of typical military HF or VHF bands for two reasons. First, we aim to corroborate these results by measurements on a real prototype setup in our future work, for which we will need to use some unlicensed band. Second, the ISM band has actually become relevant for armed forces nowadays, because adversaries are using cheap off-the-shelf radio transceivers to operate unmanned aerial vehicles (UAVs), or even toy multicopters, and improvised explosive devices (IEDs).

For the numerical results, the red transmitter's power used for controlling the UAV or IED is set to 17 dBm, while the blue team is using a transmit power of 20 dBm for both communication and jamming. The path loss at distance *d* (in kilometers) is modeled as $125 + 36 \cdot \log_{10}(d)$ [dB], which roughly represents urban Hata propagation at the ISM band with typical antenna height. Furthermore, the noise floor in all the receivers is assumed to be –90 dBm. These assumptions allow us to determine receiver SINRs based on link budget calculations given the radios' positions.

Figure 5 illustrates the red receiver's signal quality when it is located in different positions while the other transceivers are located at the coordinates shown in the figures. For reference, $\mathrm{SNR}_{bb} \approx 21$ dB, while corresponding SINR_{bb} depends on the residual SI level that would be achieved in practice. The upper part of each plot shows signal quality in the red receiver when the blue receiver is using its STAR capability for transmitting jamming, while the lower part shows the corresponding reference case without jamming.

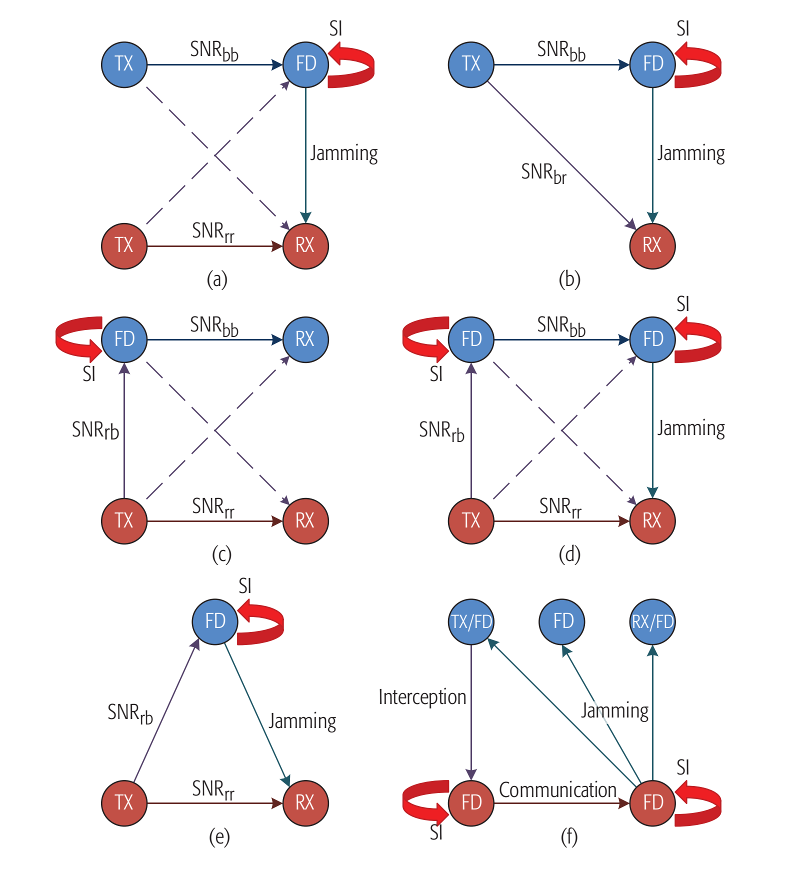


Figure 4. Military applications of the STAR capability of FD radios at the cost of SI: a) simultaneous communication and jamming against communication; b) simultaneous communication and jamming against interception; c) simultaneous communication interception and communication; d) communication with simultaneous interception and jamming; e) simultaneous interception and jamming against communication; f) counterparts with FD radios in both teams. Solid lines denote intended signals, while dashed lines represent unintended co-channel interference.

In principle, the lighter yellow color indicates better signal quality for the red team, while the signal level is below noise and jamming interference in the dark green region. We see that jamming in the FD radio act as as a "radio shield" preventing the red team from controlling the UAV, detonating the IED, or intercepting the blue transmission in the vicinity of the blue receiver.

Conclusions

Extrapolating from the rapid advances in civilian/commercial FD radios, we believe that the disruptive and unprejudiced idea of inband STAR operation also finds its way in some form to the field of military communications sooner or later. We may even be witnessing the beginning of a paradigm shift in tactical communications and electronic warfare at the moment. Thus, this article explores the prospects of FD technology in cyber-electromagnetic battles in order to inspire more scientific research on this emerging topic and to disseminate the idea within the military communications community. It is not out of the question that armed forces could gain a major technical advantage over an opponent that does not possess FD technology, or that they need new communication procedures and tactics to

It is not out of the question that armed forces could gain a major technical advantage over an opponent that does not possess FD technology, or that they need new communication procedures and tactics to counteract opponents' STAR capability. There is much room for original research in this area.

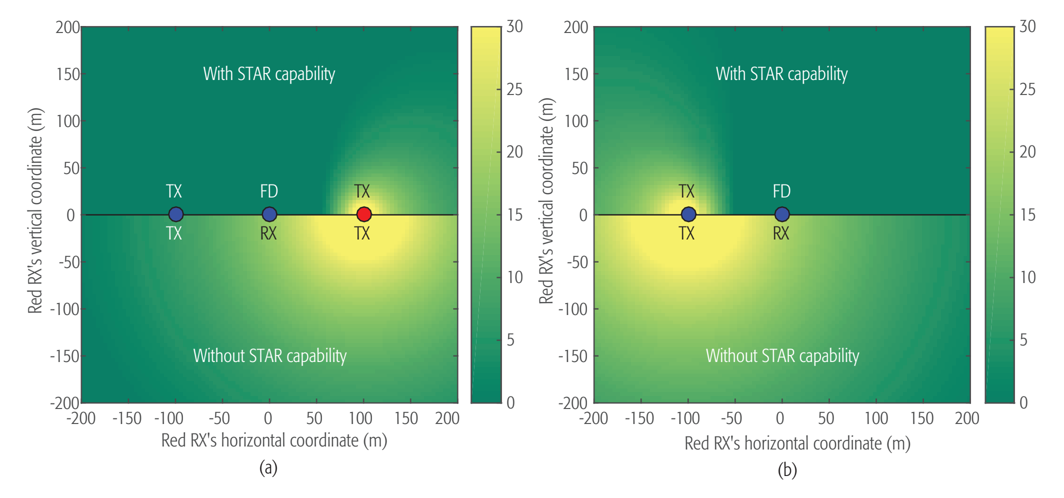


Figure 5. Comparison of HD and FD systems: a) SNR_{rr} (dB) without STAR capability and SINR_{rr} (dB) in simultaneous communication and jamming against communication; b) SNR_{br} (dB) without STAR capability and SINR_{br} (dB) in simultaneous communication and jamming against interception.

counteract opponents' STAR capability. In conclusion, we see that there is much room for original research in this area.

Acknowledgment

The research work leading to this article was supported by the Finnish Scientific Advisory Board for Defence (MATINE — Maanpuolustuksen tieteellinen neuvottelukunta) under the project "Full-Duplex Radio Technology in Military Applications" ("Full-duplex radioteknologia sotilaskäytössä").

References

[1] A. Sabharwal *et al.*, "In-Band Full-Duplex Wireless: Challenges and Opportunities," *IEEE JSAC*, vol. 32, no. 9, Sept. 2014, pp. 1637–52.
[2] Z. Zhang *et al.*, "Full-Duplex Wireless Communications: Challenges, Solutions and Future Research Directions," *Proc. IEEE*, vol. 104, no. 7, July 2016, pp. 1369–1409.
[3] S. Hong *et al.*, "Applications of Self-Interference Cancellation in 5G and Beyond," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 114–21.
[4] M. Heino *et al.*, "Recent Advances in Antenna Design and Interference Cancellation Algorithms for In-Band Full Duplex Relays," *IEEE Commun. Mag.*, vol. 53, no. 5, May 2015, pp. 91–101.
[5] T. Riihonen *et al.*, "Mitigation of Loopback Self-Interference in Full-Duplex MIMO Relays," *IEEE Trans. Signal Processing*, vol. 59, no. 12, Dec. 2011, pp. 5983–93.
[6] M. Duarte *et al.*, "Experiment-Driven Characterization of Full-Duplex Wireless Systems," *IEEE Trans. Wireless Commun.*, vol. 11, no. 12, Dec. 2012, pp. 4296–4307.
[7] T. Riihonen *et al.*, "On the Prospects of Full-Duplex Military Radios," *Proc. Int'l. Conf. Military Commun. and Info. Sys.*, Oulu, Finland, May 2017.
[8] G. Zheng *et al.*, "Improving Physical Layer Secrecy Using Full-Duplex Jamming Receivers," *IEEE Trans. Signal Processing*, vol. 61, no. 20, Oct. 2013, pp. 4962–74.
[9] D. Korpi *et al.*, "Full-Duplex Mobile Device: Pushing the Limits," *IEEE Commun. Mag.*, vol. 54, no. 9, Sept. 2016, pp. 80–87.
[10] N. Suri *et al.*, "Communications Middleware for Tactical Environments: Observations, Experiences, and Lessons Learned," *IEEE Commun. Mag.*, vol. 47, no. 10, Oct. 2009, pp. 56–63.
[11] S. L. Cotton *et al.*, "Millimeter-Wave Soldier-to-Soldier Communications for Covert Battlefield Operations," *IEEE Commun. Mag.*, vol. 47, no. 10, Oct. 2009, pp. 72–81.
[12] N. Suri *et al.*, "Peer-to-Peer Communications for Tactical Environments: Observations, Requirements, and Experiences," *IEEE Commun. Mag.*, vol. 48, no. 10, Oct. 2010, pp. 60–69.
[13] A. E. Spezio, "Electronic Warfare Systems," *IEEE Trans. Microwave Theory Tech.*, vol. 50, no. 3, Mar. 2002, pp. 633–44.
[14] G. F. Elmasry, "The Progress of Tactical Radios from Legacy Systems to Cognitive Radios," *IEEE Commun. Mag.*, vol. 51, no. 10, Oct. 2013, pp. 50–56.
[15] J. Mietzner *et al.*, "Responsive Communications Jamming Against Radio-Controlled Improvised Explosive Devices," *IEEE Commun. Mag.*, vol. 50, no. 10, Oct. 2012, pp. 38–46.

Biographies

Taneli Riihonen [S'06, M'14] (taneli.riihonen@iki.fi) received his D.Sc. degree in electrical engineering (with honors) from Aalto University, Finland, in 2014. Since 2005, he has held various research positions in the Department of Signal Processing and Acoustics, Aalto University School of Electrical Engineering. In 2015, he worked as a visiting scientist and an adjunct assistant professor at Columbia University, New York. His research activity is focused on physical-layer OFDM(A), multiantenna, relaying, and full-duplex wireless techniques.

Dani Korpi [S'14] (dani.korpi@tut.fi) received his B.Sc. and M.Sc. degrees (both with honors) in communications engineering from Tampere University of Technology, Finland, in 2012 and 2014, respectively. He is currently a researcher in the Laboratory of Electronics and Communications Engineering at the same university, pursuing his D.Sc. (Tech.) degree in communications engineering. His main research interest is the study and development of inband full-duplex radios, with a focus on analyzing the RF impairments.

Olli Rantula (olli.rantula@aalto.fi) received his B.Sc. and M.Sc. degrees (both with honors) in electrical engineering from Aalto University in 2015 and 2017, respectively. He is currently a research student in the Department of Signal Processing and Acoustics, Aalto University School of Electrical Engineering. His main research interests are focused on signal processing for medical applications.

Heikki Rantanen (heikki.rantanen@mil.fi) received his M.Sc. degree in electrical engineering from Tampere University of Technology in 1986. Since 1987, he has worked in the Finnish Defence Forces in different positions. In 2006 and 2007, he was a seconded national expert in the European Defence Agency. He is currently a principal scientist in the Information Technology Division of the Finnish Defence Research Agency. His main research interests include tactical radio communications and software-defined radio technology.

Tapio Saarelainen (tapio.saarelainen@mil.fi) received his Ph.D. degree in military sciences from the Finnish National Defence University in 2013. He has held different positions in the Finnish Defence Forces since 1990, currently serving as a staff officer in the Research and Development Division of the Army Academy. He is also an adjunct professor of applied electronics at Lappeenranta University of Technology, Finland. His research interests include adaptive sensor systems, unmanned aerial vehicles, and tactical communications.

Mikko Valkama [S'00, M'02, SM'15] (mikko.e.valkama@tut.fi) received his M.Sc. and D.Sc. degrees (both with honors) from Tampere University of Technology in 2000 and 2001, respectively. In 2003, he worked as a visiting researcher at San Diego State University, California. Currently, he is a full professor and head of the Laboratory of Electronics and Communications Engineering at Tampere University of Technology. His research interests include communications signal processing, cognitive radio, full-duplex radio, radio localization, and 5G systems.

Data Leakage Prevention for Secure Cross-Domain Information Exchange

Kyrre Wahl Kongsgård, Nils Agne Nordbotten, Federico Mancini, Raymond Haakseth, and Paal E. Engelstad

Abstract

Cross-domain information exchange is an increasingly important capability for conducting efficient and secure operations, both within coalitions and within single nations. A data guard is a common cross-domain sharing solution that inspects the security labels of exported data objects and validates that they are such that they can be released according to policy. While we see that guard solutions can be implemented with high assurance, we find that obtaining an equivalent level of assurance in the correctness of the security labels easily becomes a hard problem in practical scenarios. Thus, a weakness of the guard-based solution is that there is often limited assurance in the correctness of the security labels. To mitigate this, guards make use of content checkers such as dirty word lists as a means of detecting mislabeled data. To improve the overall security of such cross-domain solutions, we investigate more advanced content checkers based on the use of machine learning. Instead of relying on manually specified dirty word lists, we can build data-driven methods that automatically infer the words associated with classified content. However, care must be taken when constructing and deploying these methods as naive implementations are vulnerable to manipulation attacks. In order to provide a better context for performing classification, we monitor the incoming information flow and use the audit trail to construct controlled environments. The usefulness of this deployment scheme is demonstrated using a real collection of classified and unclassified documents.

Introduction

The need for efficient information exchange within national armed forces and coalitions, and between military and civilian entities has received significant attention in recent years. This need is in strong contrast to the traditional approach to securing classified military systems, where isolation of security domains and information systems has been the default approach. Thus, concepts such as NATO's information exchange gateways (IEGs) and similar initiatives within nations have been established to enable cross-domain information exchange in a secure manner.

These cross-domain solutions are required to perform various security controls (e.g., information flow control, antivirus, and access control) to ensure that the interconnection does not compromise confidentiality, integrity, or availability. In addition, non-security-specific requirements such as what type of information needs to be exchanged (e.g., friendly force identification, chat, or documents) and protocol-specific details may also impact security and what type of security controls are required. A key challenge, particularly when interconnecting domains at different classification levels, is to ensure sufficient assurance in the information flow control so that classified data is not leaked.

To improve the overall security of cross-domain solutions, the authors investigate more advanced content checkers based on the use of machine learning. Instead of relying on manually specified dirty word lists, we can build data-driven methods that automatically infer the words associated with classified content. However, care must be taken when constructing and deploying these methods as naive implementations are vulnerable to manipulation attacks.

Solutions for collaboration and information sharing across security domains may generally be categorized as transfer solutions or access solutions. A transfer solution enables the transfer of information from one domain to another, while an access solution provides a user access to services and/or information within another domain without logically transferring the information from that domain. In the latter case the access solution itself may be viewed as an extension of the domain to be accessed, imposing the domain separation requirements on the access solution (e.g., a thin client connected by a secure tunnel). Transfer solutions may be further categorized based on their ability to provide one-way or two-way transfer. For example, one-way data diodes are frequently used when information needs to be moved from a lower classified domain to a higher classified domain, while two-way information exchange may be enabled using a security filter or guard. Here we use the term guard to refer to solutions basing their release decisions (at least partly) on security labels, while it may otherwise perform similar checks as a security filter (e.g., ensuring that data objects are according to some predefined format).

Assuming that security labels are correct, a guard may provide stronger security than a security filter alone, as a security filter typically may be bypassed by anyone knowing the allowed message format. This may to some extent be mitigated by having the security filter authenticate senders, but the use of security labels nevertheless provides an additional layer of security and also better applies to content whose sensitivity typically cannot be determined by its format or type, such as documents, emails, and chat messages.

Before a user or service can initiate a request

The authors are with the Norwegian Defence Research Establishment.

Digital Object Identifier: 10.1109/MCOM.2017.1700235

0163-6804/17/$25.00 © 2017 IEEE

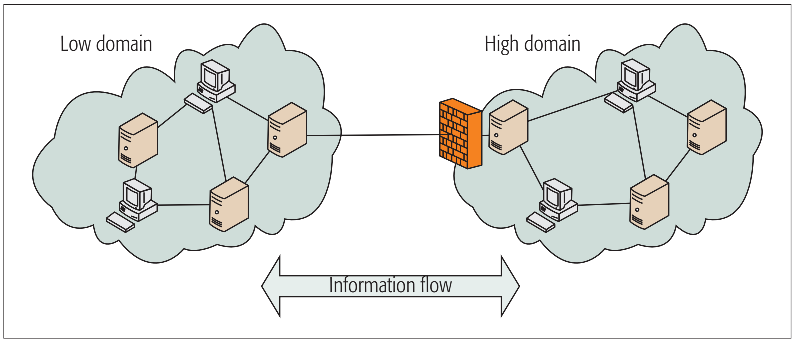


Figure 1. The data guard enables two-way information flow between “high” and “low” domains. Each object passing through the guard will have its security label validated and its content checked according to policy/configuration (e.g., content may be scanned for malware and the presence of “dirty words”), and its sender and destination fields may be verified and subject to access control. Having passed these checks, the object is then released on the condition that its security label is such that it is considered, as specified by the governing security policy, to be releasable.

to export a data object, it must first be assigned a security label. This label is cryptographically bound to the data object. While the integrity of the data object and security label as such is cryptographically protected during transfer and storage, it is much more difficult to ensure that the correct security label is attached in the first place. For instance, if a RESTRICTED document is labeled as UNCLASSIFIED, it may result in it being released to an unclassified environment (i.e., leaked). Such mislabeling may be due to human or technical errors, or users or malware trying to bypass security controls.

While the use of high assurance operating systems and applications may significantly reduce the risk of technical errors and malware, the use of commodity general-purpose operating systems and applications are often mandated due to practical and economic reasons. This lack of assurance in end-user systems may in some cases be mitigated by labeling data objects based on origin, where a potentially high assurance intermediary mechanism (e.g., gateway) labels all data from a given origin (e.g., computer or network) with a given classification (e.g., RESTRICTED). However, this approach would not allow documents from the same origin to have different security classifications. Thus, while applicable in some scenarios, this approach is often too inflexible to be practical. In the more general cases, the security label needs to be determined based on the content, rather than the origin, of the data object.

To mitigate the risk of incorrect security labels, another layer of protection in the form of a *content checker* may be applied. For text-based data objects a “dirty word list” is often used, which scans the object for the presence of keywords that are often associated with classified content, for example, security classifications, certain technical terms, locations, and project acronyms. The effectiveness of these content checkers is fully dependent on the quality of the rather static dirty word list in use. Given more recent advances in the use of machine learning, data-driven content checkers based on machine learning have the potential to improve security of guard-based cross-domain solutions.

This article highlights our experiences in developing secure, scalable, and robust cross-domain solutions (using data guards — Fig. 1) and methods for increasing the assurance in the correctness of the user or application assigned security labels. Furthermore, it provides an in-depth view into the security challenges faced when using machine learning to create data-driven content checkers for data leakage prevention (DLP).

The remainder of the article is organized as follows. The following section discusses the design and architecture of two guard prototype implementations for handling the exchange of SOAP and Extensible Messaging and Presence Protocol (XMPP) messages, showing how such solutions can be realized with high assurance. Then we discuss how machine learning methods can be utilized for DLP in a cross-domain setting to mitigate the limited assurance in the correctness of security labels, the security concerns that must be addressed when doing so, as well as engineering aspects such as which features to include and how to train and deploy said methods in order to maximize both the security and performance of the end system. Experiments regarding the detection of data leaks are then conducted and the results discussed. We then survey related work, while the conclusion summarizes the article and highlights our contributions.

Prototype High Assurance Guard

In cooperation with Thales Norway AS, we have developed two prototype guard implementations, the first for use in service-oriented architecture (SOA) and the other to support cross-domain chat. The first guard [1] supports SOAP, which is an XML-based protocol for machine-to-machine communication, messages as used in web services, while the chat guard [2] supports instant messaging through XMPP. Both guards are based on the core of a military messaging guard being developed by Thales Norway and target a Common Criteria EAL 5 certification. The guards are in alignment with the HAAG protection profile proposal [3], and uses the proposed STANAG 4774 for XML confidentiality label and STANAG 4778 for binding label and data.

Fundamental to the guard design is the use of a high assurance separation kernel. While many different guard implementations exist, most of these are based on medium assurance operating systems, effectively preventing evaluation at higher assurance levels. The separation kernel ensures that different partitions (e.g., virtual machines or processes) cannot influence each other except by using well defined interaction mechanisms. This allows security-critical functions to be separated and protected from non-critical functionality and helps ensure least privilege and non-bypassability. Together, the strong separation, high assurance, and ability to control communication between components (i.e., partitions) makes for a good environment to build high assurance systems.

Functionally, the guard is separated into several different components, each implemented as one or more partitions. Central to the design is the core component, which ensures that each object passed to the guard is processed correctly. This includes subjecting the object to label and signature checks, content checking, and other configured access controls. Content checking is done through a separate component that provides a generic plug-in interface for content checkers. Depending on the scenario, different content checkers (e.g., malware scanning and/or format checking such as XML schema validation) can be included as needed. This architecture allows new content checkers to be added without risk of compromising other guard components.

Protocol adapters provide the interface toward the interconnected domains. Different protocol adapters are used to handle the specifics of a given protocol (e.g., XMPP or SOAP/HTTP). The main task of a protocol adapter is to extract protocol-dependent attributes and transform these to protocol-independent attributes used by the core component. Additional components are used for handling configuration and the public key infrastructure. This architecture makes it easier to add new protocols without changing the security-critical code of the core, and thus simplifies the certification process.

The guards' primary release control mechanism is label checking. A mandatory access control (MAC) policy specifies the label ranges allowed to pass and how to handle unlabeled or incorrectly labeled objects. Other controls are also available for configuration, including allowed source and destination addresses, integrity control, and content checkers. When multiple such controls are in effect, a message may be blocked from release if failing any one of these checks.

To support different applications and information exchange requirements, guards will have to handle messages and protocols of varying complexity. The functional needs must always be balanced against the need for security protection. An example of this is presence status, which in XMPP chat provides information about who is logged on and available. This information is very useful for users, but can also be highly sensitive since it can reveal who is on duty and allow mapping of work schedules. Whether or not to allow this information to flow between security domains depends on the scenario and the level of risk acceptance. Support for this is given through configuration of the guard. Messages may also have different types of attachments, which may pose their own security risks and require separate content checkers. Again, what is to be allowed needs to be determined by weighing the operational gain against the additional security risk.

The prototype guards are designed with high assurance certification in mind, and the risk of information leakage due to compromise or malfunctioning of the guard is thus minimized. However, the trustworthiness of the primary release control mechanism, label control, is limited by the trustworthiness of the security labels themselves.

As of now, we do not have tools that can autonomously estimate the sensitivity of a text with a degree of confidence high enough to reduce the risk of information leakage to an acceptable level. Simple dirty word lists are quite limited. It is then natural to investigate the possibility of developing more effective automated classification techniques by using recent advancements in the field of machine learning. The second part of the article is devoted to the latest research on this topic.

Machine-Learning-Based Content Checking

Machine learning lends itself naturally to the problem of classifying unstructured textual information. However, much of the available research focuses on classification of text into a set of predefined topics, which is not directly applicable to our problem. The sensitivity of a text depends not only on what it talks about, but also on the context in which it was produced and what kind of damage the information can do if leaked. This type of assessment is difficult even for a person, let alone for an algorithm.

As long as existing data has a known classification, it is possible to verify that it is not labeled incorrectly by employing direct comparison techniques like hashing. Estimating the sensitivity for completely new [4–6] or heavily processed (rewritten, summarized, etc.) [7] information, on the other hand, is more challenging and is better handled using machine learning. When the algorithm is presented examples of known classified and unclassified documents, it will attempt to infer which features are associated with each of the target classes (classification levels).

In order to further improve the performance, we worked on two ideas. The first consists of training the algorithm with an even more specific context to increase the accuracy rate. The other explores the possibility of improving the probability of detecting users (or other entities) with an abnormal amount of likely misclassifications over time, rather than aiming at detecting misclassification of single documents. In a guard setting, automatic classification may also be used to prioritize which documents are to be subject to manual review, in which case a somewhat lower classification accuracy may be acceptable.

In the remainder of this section, we provide an overview of the challenges faced and the state of the art in developing and engineering machine-learning-based content checkers for the cross-domain information exchange setting.

Features

Before any learning can take place, the documents must be transformed into feature vectors. Feature engineering refers to the process of capturing an important characteristic of a document as a numerical value (feature). It is the part of the machine learning process that requires the most in-depth domain expertise, and is, together with the size/quality of the training dataset and the choice of model class, what has the greatest impact on the performance of the resulting model.[1]

Features for textual content are primarily derived from variations of word counts/frequencies, but one also uses more general features such as the average sentence length, the number of capitalized words, and statistics regarding punctuation. Advanced features such as the part-of-speech (PoS) and named-entity recognition (NER) tags of words in a sentence are also beneficial for certain classes of tasks. A list of the features that we have used for the machine-learning-based content checker are listed below.

***n*-gram:** An *n*-gram is a contiguous sequence of *n* words. Term-frequency inverse document-frequency weights (TF-IDFs) modify these

To support different applications and information exchange requirements, guards will have to handle messages and protocols of varying complexity. The functional needs must always be balanced against the need for security protection. Examples of this includes presence status, which in XMPP chat provides information about who is logged on and available.

[1] In the field of representation learning it has been demonstrated that one can, for some tasks, with a sufficiently large dataset and the right network architecture, automate the feature engineering phase. This is one of the reasons behind the resurgence and popularity of research into deep learning methods.

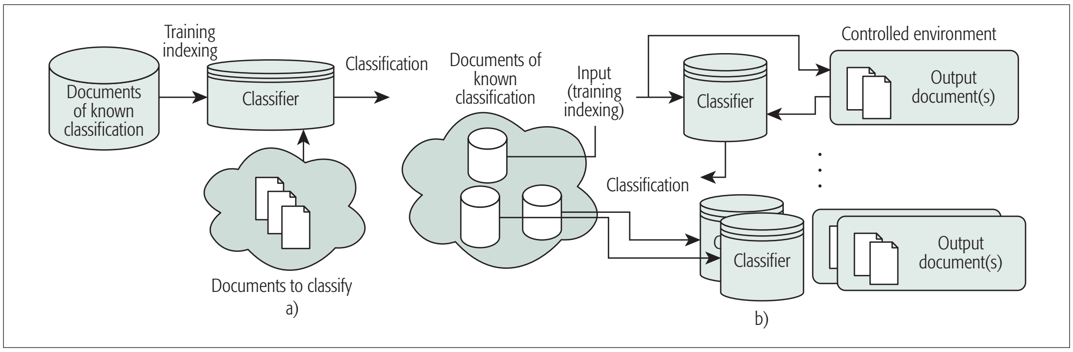


Figure 2. a) Usually, a classifier used in DLP is trained on all available documents; b) with a controlled environment, only the documents of known classification accessed from the environment are used to train the classifier, which in turn is used to classify documents generated within the environment. Multiple controlled environments can exist simultaneously, each characterized by its own input and output.

In order to assess how well this methodology performs on short (e.g., message-like) documents, we use DNSA documents as the input/training data and evaluate it on sentence(s) sampled from the corresponding abstract. For comparison we also train global classifiers that use all the available data as a training set.

frequencies to better reflect the importance of a particular *n*-gram for the document. In the bag-of-words (BoW) model a document is represented as a multiset of its *n*-grams. While the BoW model discounts word order (except for what is captured within the *n*-gram) and any grammar, it retains the semantic aspects and has been shown, despite its simplicity, to be very useful for text classification and in information retrieval systems [8]. *n*-gram frequencies can also be computed on the character level.

Lemmatization: Lemmatization is the process of grouping together the inflected forms of words; for example, "flies" is mapped to "fly" and "better" is mapped to "good." This pre-processing step could be beneficial for sparser documents and for detecting paraphrased content and the use of synonyms. Features can be extracted in parallel and concatenated into a single high-dimensional vector representation.

Controlled Environments

In a cross-domain scenario each data access request in the "high" domain can be logged on a per-user/session level. The audit trail of access requests, or the incoming information flow, can be used to derive what we have named controlled environments. A controlled environment refers to any environment where we have control on all imported documents and their respective security classifications. The set of imported documents (e.g., those accessed by the user during a session) is defined as *input*, and any new document(s) generated within the controlled environment is defined as *output*. By using the set of input documents as basis for the classifier, thereby reducing the noise in the classification process as the input documents are more relevant to the output, we can more accurately estimate the classification of output documents [7]. Our proposed solution inspects the information flow to the controlled environment as shown in Fig. 2b, and estimates the classification of output documents based on the information about the input documents.

Experiments: We want to analyze the performance for message-like (i.e., short) documents and using both a controlled environment setting as well as a traditional global classifier (one that uses the complete set of documents for training).

As a dataset we use a subset of de-classified documents from the Digital National Security Archive (DNSA). From this repository we extracted the three sub-collections:

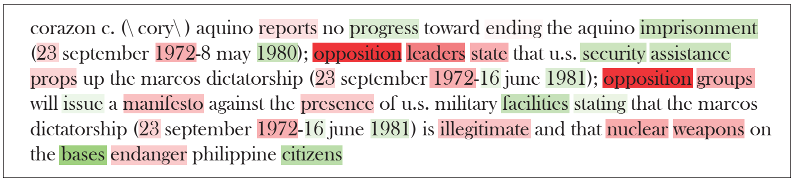
- *Afghanistan: The Making of U.S. Policy, 1973–1990*
- *China and the United States: From Hostility to Engagement 1960–1998*
- *The Philippines: U.S. Policy during the Marcos Years, 1965–1986*

These were chosen because they contained a mix of both classified and unclassified documents from unrelated domains and from partially overlapping time periods.

We train the classifiers (l_2-regularized logistic regression) on documents from the DNSA dataset [7] that were imported into a controlled environment, and then evaluate the performance on the corresponding abstracts (these are removed from the input documents). This procedure simulates how (potentially classified) information is transformed into new documents. We have also studied other transformation models, for example, the mixing of documents and the use of synonym (phrases), but we omit them from further discussions as the "abstract" transformation is the most realistic and challenging one. The leakage of known unmodified documents can be detected with very high accuracy (0.99) using both methods, and is not discussed further as this can be handled using existing methods (e.g., hashing).

In order to assess how well this methodology performs on short (e.g., message-like) documents, we use DNSA documents as the input/training data and evaluate it on sentence(s) sampled from the corresponding abstract. For comparison we also train global classifiers that use all the available data as a training set. In both cases we use the logistic regression implementation provided by the Python machine learning library scikit-learn [9]. Cross-validation (5-fold) with a randomized hyperparameter search is used to determine the optimal value for the regularization coefficient, while features were extracted using the tool `TfidfTransformer` (tf-idf weights) from the scikit-learn package and the UDPipe pipeline toolkit (lemmatization) [10].

Figure 3 shows a visualization of how the classifier analyzes a document to determine its sensitivity level. The words highlighted in red indicate terms that are associated with the more sensitive class. For the particular example shown, it is clear that the model has learned that words such as



corazon c. (\ cory\ ) aquino reports no progress toward ending the aquino imprisonment (23 september 1972-8 may 1980); opposition leaders state that u.s. security assistance props up the marcos dictatorship (23 september 1972-16 june 1981); opposition groups will issue a manifesto against the presence of u.s. military facilities stating that the marcos dictatorship (23 september 1972-16 june 1981) is illegitimate and that nuclear weapons on the bases endanger philippine citizens

Figure 3. Words highlighted in red are those associated with the classified content class, while green words are those associated with the unclassified content class. The darker the color, the stronger the signal or the connection between the feature and the class.

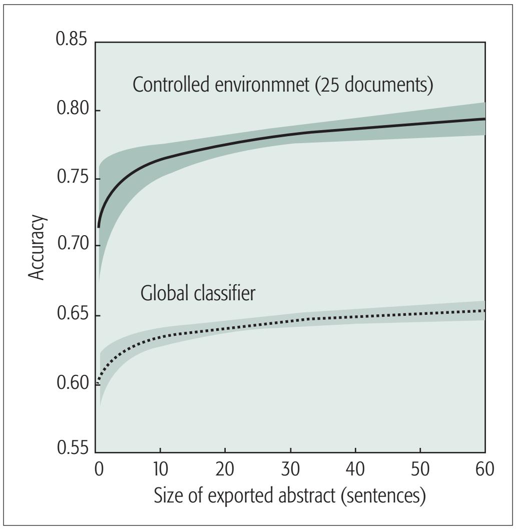


Figure 4. Content checker performance. Accuracy as a function of the number of sentences in the exported abstract sample, for both a controlled environment of size 25 and a global classifier using data derived from the DNSA dataset.

“opposition,” “nuclear,” and “endanger” are often linked with sensitive information, while the terms “progress,” “citizen,” and “imprisonment,” on the other hand, are more likely to signify a non-classified document.

A comparison between a controlled environment and a global deployment scenario is presented in Fig. 4. It shows the accuracy of the model as a function of the number of sentences from the abstract that is sent as a message. Comparing the two graphs depicted, we see that we are able to achieve a significant boost in performance when using the per-user trained (i.e., controlled environment) model instead of the traditional global classifier. While this is a surprising result, and one that seemingly contradicts the conventional wisdom that more data always provides higher accuracy, it reflects that determining sensitive content is very context-dependent and that we are exploiting the assumption that any classified content can be traced back to information contained in the imported documents. As such, a controlled environment would likely result in severe performance degradation if we wanted to use it to detect classified information that is completely unrelated to the input documents.

Internal Threat Scores

As the content checkers are plagued with nontrivial false positive rates, we have also investigated the idea of constructing a meta-score called the *internal/insider threat score* (ITS), which uses the aggregated confidence scores to detect long-term discrepancies between the user-assigned sensitivity level and the sensitivity level predicted by the machine learning model [11]. It works by modeling how the users (or other entities) assign labels as a generative process and then infers (using a Bayesian network model) the latent variables that describe how often documents are misclassified in general for each user. These misclassification rates are what we use to compute the ITS. On a more technical level, the model captures to what extent the deviations of the confidence score distribution for one user and the confidence score distribution for known classified documents can be attributed to incorrectly assigned labels by the user. By operating on a per-user (as opposed to per-document) level, the number of false alarms is reduced. Figure 5 displays a visualization of the ITS.

Another concern that must be analyzed and addressed is the threat of the content checker itself becoming a target for an attacker.

Secure Machine Learning

A core underlying principle behind most machine learning algorithms and tasks is that the training and evaluation datasets are generated from the same unknown distribution, that is, it assumes a stationary environment. Under this assumption, minimizing the empirical risk (informally the error) on the smaller training dataset, which has often been painstakingly hand labeled, is equivalent to minimizing the risk on the larger evaluation dataset. However, this assumption is violated for security tasks such as intrusion detection and DLP systems, where one must take into account the possibility of attackers actively seeking to bypass detection by manipulating the classifier itself. A machine learning algorithm is said to be secure if it performs adequately when deployed in adversarial conditions.

Security assessments of machine learning systems are conducted with respect to the three axes [12]:

Influence: A user can influence the learning system by conducting either a causative or an exploratory attack. Causative (interchangeably: poisoning) refers to manipulating (parts of) the training data with the intention of exerting control of the learning process. Exploratory refers to inducing and exploiting a misclassification, for example, by rewriting a classified document such as not to trigger the content checker.

Security Violation: Security violations take on one of two forms: integrity, for example, sensitive content being incorrectly classified and let through the guard, and availability, for example, non-sensitive data being misclassified en masse, which may effectively render the system useless.

Specificity: The scope of the attack can be either a targeted or an indiscriminate attack. An attack of particular concern in a content checker context is the possibility of data leaking from the model itself. If a user knows the functional form of the classifier (e.g., whether we are using a logistic regression, support vector machines, or some other model), and if the user can probe it for numerical outputs, that is, if he has access to the probability/confidence scores for each data point, through repeated experiments he can recover the

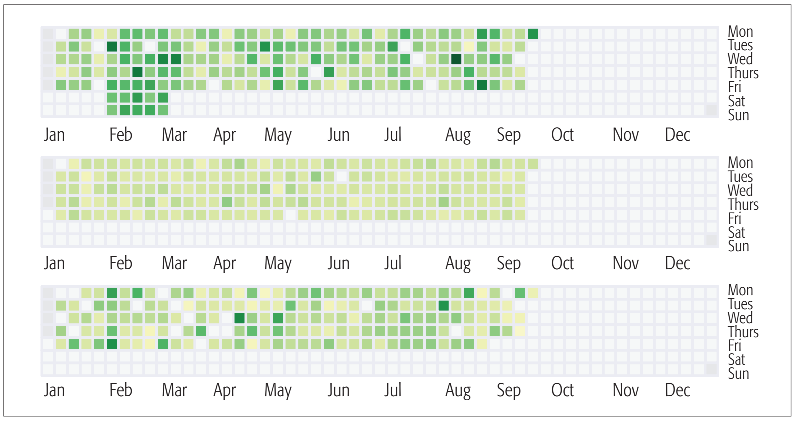


Figure 5. A heatmap time series visualization of the daily ITS value (misclassification rate) for three users during a nine- month simulation period. A darker shade of green signifies a higher ITS value. Top: A malicious user that has a very high baseline misclassification rate and periods of increased weekend activity; middle: a regular user with a low misclassification rate; bottom: an incompetent user with a high misclassification rate.

actual parameter values of the underlying model or (parts of) data points in the training set [13]. When the training set contains classified information, one must be particular wary of the possibility of the model leaking data.

We can analyze the security risks for the inferred model with respect to the attack categories/classes, estimate the feasibility of said methods, as measured in terms of the cost (risk and resources) incurred by the attacker, and propose potential mitigation steps.

Exploratory (Insider Attacker): A malicious user can, in theory, always bypass the detection mechanism by rewriting a document such as not to trigger the alarm. While this procedure can be automated for images [14], it remains a manual process for unstructured text. Taken to its extreme, we arrive at a scenario in which the insider employs methods of steganography to covertly embed classified information within other innocuous content. There is no generic solution that solves this, and any system must be combined with host-based systems to detect the presence of stenographic software.

Causative (Insider Attacker, Controlled Environment): By carefully choosing what to import, the training set can potentially be shaped in such a way that the algorithm later misclassifies documents containing classified content that the user wants to exfiltrate. Defenses against these attacks include algorithms that effectively sanitize the data by modifying the learning process to dynamically discount those data points in the training set that have a significant negative impact on the performance. A competing class of defense mechanisms recasts the problem as one of anomaly detection, for example, do the model parameters of one user deviate dramatically from the model parameters of other users. Similarly, performing a causative attack to exfiltrate larger amounts of data would likely result in detectable anomalies in the set of imported documents for a controlled environment setting.

Instance-based algorithms (e.g., k-NN) are not as susceptible to causative attacks because there is no training phase involved, and we can use a decision strategy in which any imported document with a similarity score greater than a threshold value will result in assigning the strictest label (e.g., "Classified") to the document.

Model Data Leakage: When the model is invoked by the trusted guard, there is no known way of performing such an attack as the user does not have access to the confidence scores. Furthermore, with a controlled environment the user is already authorized to access the documents in the training set, which renders such an attack meaningless.

Related Work

The first work studying the use of machine learning to predict the security classification level of textual documents was done by Brown *et al*. [4], who built a binary classifier using only the abstract section of documents. Similar studies later reproduced and expanded upon these results [5, 6, 15] by using the complete document contents, multiple security classification levels, and per-paragraph sensitivity predictions, while the concept of controlled environments was introduced by us in [7].

Conclusion

In this article we have discussed the use of high assurance data guards for controlling the information flow between different security domains.

We observe that one is currently able to develop guards whose assurance level in practical scenarios surpasses the assurance provided by the security labels on which the guard relies. Thus, while the guard only releases data objects with a security label releasable by policy, there is typically less confidence that those security labels are correct. Thus, more effective content checkers that detect such mislabeled data objects would be of significant value.

We have introduced the concept of applying machine learning techniques to construct automated, data-driven content checkers. Our treatment of the topic extends beyond the theoretical considerations by including what we have, through extensive experiments, observed to be the best practices for data-driven content checkers, including which features to use and how to deploy the classifiers. We have focused especially on assessing the impact different deployment settings have on the security and performance of the end system. As such, we have presented the concept of controlled environments, where the audit trail or incoming information flow is used to construct per-user/session classifiers, which yielded a significant improvement in performance. The proposed methods have also been analyzed with respect to causative, exploratory, and data leakage attacks, and we noted that while they still remain vulnerable to causative attacks carried out by sophisticated insiders, they completely alleviate the threat of the inferred model leaking sensitive data from the training set.

While previous work looked at building classifiers for complete documents [7], we have extended these methods to work for shorter messages, which is a more difficult case.

The performance we currently achieve is not sufficient to warrant a fully automated deployment scheme. However, with an appropriate decision threshold, the classifiers can be used to determine which documents must be manually assessed. A meta-score (ITS) operating on the per-user long-term classifications trends can also be used to further reduce and manage the number of false alarms [11]. As future work, one can also investigate the feasibility and usefulness of other forms of classifiers, such as language and genre detection for increasing the trust in exported documents. We have conducted preliminary experiments using an authorship verification model, built using the stylometric information embedded in the past chat messages of users, to detect instances in which an outgoing message was not authored by the user in question. Combining the results from such different types of classifiers may potentially help improve accuracy.

While the guard only releases data objects with a security label releasable by policy, there is typically less confidence that those security labels are correct. Thus, more effective content checkers that detect such mislabeled data objects would be of significant value.

References

[1] R. Haakseth *et al.*, "A High Assurance Guard for Use in Service-Oriented Architectures," *Proc. Int'l. Conf. Military Commun. Info. Systems*, 2015.

[2] R. Hakseth *et al.*, "Cross Domain Communications Using an XMPP Chat Guard," FFI-rapport 17/01491, 2017.

[3] K. Wrona and N. Menz, "Protection Profile for the NATO High Assurance ABAC Guard (HAAG), version 1.3," NCIA tech. rep. TR-2012-SPW0084-18-13-4, 2013.

[4] J D. Brown and D. Charlebois, "Security Classification Using Automated Learning (Scale): Optimizing Statistical Natural Language Processing Techniques to Assign Security Labels to Unstructured Text," tech. rep., DTIC doc., 2010.

[5] H. Hammer *et al.*, "Automatic Security Classification by Machine Learning for Cross-Domain Information Exchange," *Proc. IEEE MILCOM*, vol. 31, 2015.

[6] K. Wrona *et al.*, "Assisted Content-Based Labelling and Classification of Documents, *Proc. 2016 Int'l. Conf. Military Commun. Info. Systems*, 2016, pp. 1–7.

[7] K. W. Kongsgård *et al.*, "Data Loss Prevention Based on Text Classification in Controlled Environments," *Proc. Info. Systems Security*, Springer, 2016, pp. 131–50.

[8] C. D. Manning *et al.*, *Introduction to Information Retrieval, Volume 1*, Cambridge Univ. Press, 2008.

[9] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in python," *J. Machine Learning Research*, 12 Oct. 2011, pp. 2825–30.

[10] M. Straka, J. Hajic, and J. Straková, "Ud-Pipe: Trainable Pipeline for Processing Conll-u Files Performing Tokenization, Morphological Analysis, PoS Tagging and Parsing," *Proc. 10th Int'l. Conf. Language Resources and Evaluation*, 2016.

[11] K. W. Kongsgård *et al.*, "An Internal/Insider Threat Score for Data Loss Prevention and Detection," *Proc. ACM Int'l. Wksp. Security and Privacy Analytics*, 2017.

[12] M. Barreno *et al.*, "The Security of Machine Learning," *Machine Learning*, vol. 81, no. 2, 2010, pp. 121–48.

[13] Z. Ji, Zachary, C Lipton, and C. Elkan, "Differential Privacy and Machine Learning: A Survey and Review," arXiv peprint arXiv:1412.7584, 2014.

[14] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in Machine Learning: From Phenomena to Blackbox Attacks using Adversarial Samples," arXiv preprint arXiv:1605.07277, 2016.

[15] K. Alzhrani *et al.*, "Automated US Diplomatic Cables Security Classification: Topic Model Pruning vs. Classification Based on Clusters," arXiv preprint arXiv:1703.02248, 2017.

Biographies

Kyrre Wahl Kongsgaard (kyrre-wahl.kongsgard@ffi.no) is a final year graduate student (Ph.D. degree) in the Institute for Technology Systems, University of Oslo, and a scientist at the Norwegian Defence Research Establishment (FFI). He received his M.Sc. degree in computational science from the University of Amsterdam in 2013. His current research is concerned primarily with applying machine learning techniques to information security problems such as DLP and network intrusion detection systems.

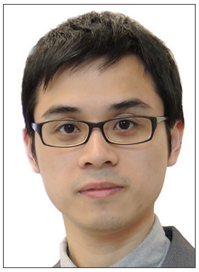
Nils Agne Nordbotten (nils.nordbotten@ffi.no) is a research manager and principal scientist within cybersecurity at FFI, and an adjunct associate professor at the University of Oslo (UiO). He received his Ph.D. (2008) and Cand.scient. (2003) in computer science from UiO, and an executive Master of Management degree (2012) from BI Norwegian Business School.

Federico Mancini (federico.mancini@ffi.no) received his M.Sc. degree in computer science in 2004 from Università degli Studi Roma Tre, Italy, and his Ph.D. degree in algorithms and graph theory in 2008 from the University of Bergen (UiB), Norway. He has since then shifted his research focus to information security and is currently a senior scientist at FFI. He was also an adjunct assistant professor at UiB from 2011 to 2015.

Raymond Haakseth (raymond.haakseth@ffi.no) is a research manager and senior scientist within ICT and cybersecurity at FFI, where he has worked since 2004. He received his M.Sc. degree in computer science from the University of Tromso in 2003. His research interests include information assurance, distributed systems, and software engineering.

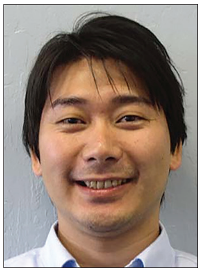
Paal E. Engelstad (paal.engelstad@ffi.no) received his Ph.D. in computer science from UiO in 2005. He is now working as a research scientist at FFI, an adjunct professor at UiO, and vice-dean and full professor at UiO and Akershus University College. His current research interests include fixed, wireless, and ad hoc networking, cybersecurity, machine learning, and distributed, autonomous systems.

IoT and Information Processing in Smart Energy Applications


Wei-Yu Chiu

Hongjian Sun

John Thompson


Kiyoshi Nakayama

Shunqing Zhang

There are several challenges for the current electricity grid: growing electricity demand, an aging grid infrastructure, ever increasing penetration of renewables, and significant uptake of electric vehicles and energy storage with behind-the-meter applications for residential and commercial buildings. To address these challenges, there must be strong and low-cost communications infrastructures that can support rapid and secure information exchange as well as consistent and efficient design of communication protocols and architectures to enable automation and effective use of smart energy resources.

The Internet of Things (IoT) could accelerate establishment of such infrastructures. With IoT technologies, many more devices could be controlled and managed through the Internet; data pertaining to the grid, commercial buildings, and residential premises can readily be collected and utilized. To derive valuable information from the data, further information and data processing become essential.

This Feature Topic aims to disseminate general ideas extracted from cutting-edge research results relevant to smart energy applications from the perspectives of IoT, and advanced information processing and communications technologies. We received 23 submissions from around the world and selected 6 papers for publication (26 percent acceptance rate). The authors of these accepted papers share various viewpoints and the latest findings from their research and ongoing projects.

For smart grid applications, we need to predict the electrical load so that the underlying smart grid can effectively balance the power supply and demand. In general, predictions are made based on the data obtained using IoT and smart meter technologies. While collecting data becomes easier to manage, analyzing it becomes more complex and time-consuming because of the 4V properties of big data (i.e., volume, velocity, variety, and veracity). Conventional methods seem to struggle with analyzing complicated data relationships. Our first article, "When Weather Matters: IoT-Based Electrical Load Forecasting for Smart Grid," addresses this difficulty. An IoT-based deep learning system is presented that can automatically extract features from captured big data, ultimately yielding accurate estimation of future loads.

To successfully realize smart energy management, we need a robust communications infrastructure for ubiquitous, large-scale, and reliable information exchange among sensors and actuators. The robustness is essential because sensors and actuators deployed in the field will often have little or no human interaction. Two relevant studies are included. The second article, "Software Defined Machine-to-Machine Communication for Smart Energy Management," discusses an interesting software-defined machine-to-machine framework. Under this framework, cost reduction, fine-granularity resource allocation, and end-to-end quality of service are considered. Several open issues and key research opportunities are identified. The third article, "5G Mobile Cellular Networks: Enabling Distributed State Estimation for Smart Grids," presents advanced distributed state estimation methods in a 5G environment. Emerging distributed state estimation solutions are provided, and their integration as part of the future 5G-based smart grid services is investigated.

We cannot emphasize enough the importance of data security when IoT technologies are employed for smart energy applications. There is an increasing need for defense mechanisms that either protect the system from attackers in advance or detect data injection attacks in real time. A comprehensive tutorial and survey regarding data security threats and associated research challenges can be found in the fourth article, "Energy Big Data Security Threats in IoT-Based Smart Grid Communications." The fifth article, "Defense Mechanisms against Data Injection Attacks in Smart Grid Networks," investigates relevant signal processing techniques and introduces an adaptive scheme for detecting malicious data injection. Another solution to secure communications is the use of power talk, a low-rate communication technique for direct current microgrids. This technique may support several smart energy applications and is reviewed in the sixth article, "Resilient and Secure Low-Rate Connectivity for Smart Energy Applications through Power Talk in DC Microgrids."

These six articles provide an excellent overview on advanced technologies for smart energy applications. We would like to thank the authors and reviewers for their contributions and support. We also hope this Feature Topic can motivate researchers in academia, practitioners in industry, and officials in government to explore more possibilities in this exciting area in the future.

Biographies

Wei-Yu Chiu [M'11] (wychiu@ee.nthu.edu.tw) received his Ph.D. degree in communications engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2010. He was a postdoctoral research fellow at Princeton University from 2011 to 2012 and a visiting scholar at Oklahoma State University in 2015. He is currently an assistant professor of electrical engineering at NTHU. His research interests include multi-objective control, smart grid, and computational intelligence.

Hongjian Sun [S'07, M'11, SM'15] received his Ph.D. degree from the University of Edinburgh, United Kingdom, in 2011, and then took postdoctoral positions at King's College London, United Kingdom, and Princeton University. Since 2013, he has been with the University of Durham, United Kingdom, as an assistant professor (2013–2017) and then an associate professor (reader). His research mainly focuses on smart grid areas: communications and networking, demand side management and demand response, and renewable energy sources integration.

John Thompson [M'94, SM'13, F'16] currently holds a Personal Chair in Signal Processing and Communications at the University of Edinburgh. He was deputy academic coordinator for the Mobile Virtual Centre of Excellence Green Radio project and now leads the U.K. SERAN project for 5G wireless. He also currently leads the European Marie Curie Training Network ADVANTAGE, which trains 13 Ph.D. students in smart grids. He was a Distinguished Lecturer on green topics for ComSoc, 2014–2015.

Kiyoshi Nakayama [M'14] completed his Ph.D. degree in computer science at the University of California, Irvine in June 2014, and then belonged to Fujitsu Laboratories of America as a postdoctoral research associate in the smart energy research group. Since 2015, he has been a researcher at NEC Laboratories America leading the microgrid and behind-the-meter project for developing and deploying a cloud-based autonomous energy management platform with intelligent data and model-driven approaches.

Shunqing Zhang [S'05, M'09, SM'14] received his Ph.D. degree in 2009 from the Department of Electrical and Computer Engineering, Hong Kong University of Science and Technology. He was a senior researcher with Huawei, and then a research scientist and deputy director of ICRI-MNC, Intel Labs. He is currently a full professor at Shanghai University, China. His current research interests include cellular network intelligence, energy-efficient 5G communication networks, and non-orthogonal waveform design.

Call for Papers

IEEE Communications Magazine

Information-Centric Networking Security

Background

Information-centric networking (ICN) is a new network architecture that provides the access to named data as first-order network service, providing better trust in data authenticity and greater potential for optimizing forwarding behavior compared to traditional host-based communication systems like the Internet today.

The ICN principle of accessing authenticated named data in the network enables several optimizations, such as network-layer data caching, flexible multipath communication and simplified mobility management. ICN thus addresses many important requirements of applications such as high-performance, scalable media data distribution and reliable distributed Internet of Things networks.

ICN is an active research area that includes specific topics such as network architectures, applications, transport, and caching techniques. Security is a particularly important topic since ICN enables new approaches with respect to confidentiality, access control and trust management that we want to address by this Feature Topic.

Solicited topics include (but are not limited to):

- Security architectures for information-centric networking (ICN)
- Authentication and authorization for distributed caching environment
- Access control mechanisms for ICN
- Security in mobile ICN
- Cryptographic protocols against internal attacks in information-centric network
- Information sharing and data protection in ICN
- Secure monitoring in ICN
- Denial of service attacks prevention in ICN
- Privacy protection in ICN
- Privacy policy framework for ICN
- Privacy in caching, naming, signature
- Mechanisms to enforce privacy and trust
- Privacy-preserving processing in ICN
- Privacy-preserving data publishing on ICN
- User privacy, data providers privacy and ICN application platform privacy
- Privacy and security in ICN applications
- Trusted computing platform for ICN
- Trust models for ICN
- Trust management for ICN
- Copyright management and business models for ICN

Submissions

Articles should be tutorial in nature, with the intended audience being all members of the communications technology community. They should be written in a tutorial 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, excluding figures, tables and captions). Figures and tables should be limited to a combined total of six. The number of archival 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 submit 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 "July 2018 / Information-Centric Networking Security" as the Feature Topic category for your submission.

Important Dates

- Manuscript Submission Deadline: November 1, 2017
- Decision Notification: March 1, 2018
- Final Manuscript Due: April 15, 2018
- Publication Date: July 2018

Guest Editors

Xiaoming Fu	Satyajayant Misra	Dirk Kutscher	Ruidong Li
Univ. of Göttingen, Germany	New Mexico State Univ., USA	Huawei Research, Germany	NICT, Japan
fu@cs.uni-goettingen.de	misra@cs.nmsu.edu	dku@dkutscher.net	lrd@nict.go.jp

When Weather Matters: IoT-Based Electrical Load Forecasting for Smart Grid

Liangzhi Li, Kaoru Ota, and Mianxiong Dong

The authors introduce an IoT-based deep learning system to automatically extract features from the captured data, and ultimately, give an accurate estimation of future load value. One significant advantage of their method is the specially designed two-step forecasting scheme, which significantly improves the forecasting precision.

Abstract

Electrical load forecasting is still a challenging open problem due to the complex and variable influences (e.g., weather and time). Although, with the recent development of IoT and smart meter technology, people have obtained the ability to record relevant information on a large scale, traditional methods struggle in analyzing such complicated relationships for their limited abilities in handling nonlinear data. In the article, we introduce an IoT-based deep learning system to automatically extract features from the captured data, and ultimately, give an accurate estimation of future load value. One significant advantage of our method is the specially designed two-step forecasting scheme, which significantly improves the forecasting precision. Also, the proposed method is able to quantitatively analyze the influences of some major factors, which is of great guiding significance to select attribute combination and deploy onboard sensors for smart grids with vast areas, variable climates, and social conventions. Simulations demonstrate that our method outperforms some existing approaches, and can be well applied in various situations.

Introduction

Smart grid, as a power system for the future, has recently received lots of attention. Although many encouraging research works have emerged in the relevant area, one grave problem remains unsolved: electrical load forecasting. An accurate estimation of future load variation is of great significance for competitive and deregulated electricity markets, where load prediction is an important guide, both for power companies and electricity consumers, to make decisions and in operations [1].

The major obstacle in load forecasting is the numerous impact factors. There are so many possible influences that it is extremely difficult to find a meaningful relationship between load variation and these factors. In fact, even the acquisition of necessary data has not been an easy case until quite recently. The emerging of smart meter infrastructures [2], efficient sensing methods [3], and Internet of Things (IoT) technologies [4] give us a chance, for the first time, to record and analyze possible influences on a large scale. With several equipped sensors, smart meters can be used to independently capture various environmental data. Also, they can obtain the shared data from IoT-enabled devices. All these data will be uploaded to the central controller. Then a massive number of data can be accumulated for further analysis. However, it is still a challenging problem to handle the data, due to the complex and variable influences, especially the diverse weather conditions. Indeed, most existing time-series forecasting approaches [5] have some limitations when applied to electrical load prediction. The classical statistical methods are criticized for their limited abilities in handling nonlinear data, and the computational intelligence methods are facing problems like inappropriate handcrafted features, limited learning capacity, inadequate learning, inaccurate estimation, and insufficient guiding significance. Although there have been several attempts based on the state-of-the-art machine learning methods, which can partially resolve these problems, their performance can be significantly improved using some ingenious design introduced in the article.

To solve these problems, we desire to utilize the state-of-the-art deep learning methods [6] to automatically extract features from the historical data and give an accurate estimation of future load value. For the sake of data collection, we implement an IoT-enabled system in an urban area of south China, as shown in Fig. 1. Smart meters are adopted to record and upload electrical and background data, with specially designed sensors and IoT-enabled devices. They are deployed in every electricity consumption unit, and share their information with the control center. Ultimately, we obtain a large dataset over seven years, from 2010 to 2016, including all conceivable factors. Then the data is used for network training and influence analysis. As IoT and smart meters both have the ability of bidirectional communication, the recommendations and decisions made by the control center can be sent to the demand side. The IoT infrastructure in Fig. 1 is an important part of the proposed system. The smart meters can connect to all the IoT-enabled sensors, gadgets, and appliances in the electricity unit (e.g., a smart home). The proposed system can deal with all these data, and based on that, perform accurate load forecasting.

The main contributions of our work include:

- We propose an IoT-enabled load forecasting system based on the state-of-the-art deep learning technologies. Compared to the traditional time-series analysis methods, the proposed method can perform accurate prediction without hand-crafted features.
- We design an ingenious two-step forecasting scheme, which forecasts the daily total consumption at first, and based on that, predicts the

Digital Object Identifier: 10.1109/MCOM.2017.1700168

The authors are with Muroran Institute of Technology.

0163-6804/17/$25.00 © 2017 IEEE

intra-day load variation. This method can significantly improve the forecasting precision, which is demonstrated in the experiment section.

- We work out an analysis method of possible influence factors. To the best of our knowledge, this is the first attempt to gain insight into the relationship between the factors and the actual load, which, we believe, will play a tremendous role in selecting attribute combination and deploying smart meters, especially for the smart grids in some countries with vast territory and varied climates.

Recent Advances in Load Forecasting

IoT-Enabled Smart Meter Systems

A smart meter is a modernized electrical device that records energy consumption and uploads it to the utility for billing and further analysis. The most cutting-edge smart meters not only have two-way communication ability, but are equipped with real-time sensors that can gather the data on relevant factors. This kind of electricity meter is a vital part of advanced metering infrastructure (AMI), a system keeping the whole smart grid connected and informed. With AMI, the utility can obtain necessary data from the client side, and push notifications and recommendations to the clients. The connected smart meter is a fundamental component of the future smart grid, and also a cornerstone of our research. Many researchers are working in this area [7].

IoT, as a hot topic in recent years, is a good approach to implement connected AMI systems. However, like any other applications using IoT technologies, the underlying network to connect smart meters must be carefully investigated. Some researchers find it necessary to clarify the exact capacities of existing wireless networks for the upcoming smart metering traffic [8]. A few preliminary conclusions have been drawn, including decreasing the communication interval and using phasor measurement units (PMUs).

In addition, an efficient distributed communication architecture has been proposed for the connection of smart meters [9], which can leverage data processing locally. Besides, with carefully selected control centers [10], the cost of deployment and communication can be significantly decreased.

Time-Series Forecasting Methods

Although time-series forecasting is a topic with a long history, it is still an open problem due to its complexity. The existing approaches are of two main types: statistical methods and computational intelligence methods.

A statistical method is an obvious and natural solution when dealing with a series of numbers, including many algorithms with different design principles [11]. There is a famous exponential smoothing method called Holt-Winters (HW). HW is a good choice when the time-series shows both trend and seasonality. Two sub-models are included in HW: the additive model for data with additive seasonality and the multiplicative model for data with multiplicative seasonality.

Although many classical statistical methods have emerged over the past few decades, they are currently disfavored due to their limited abilities in handling complicated nonlinear relationships. Computational intelligence, one of the hottest topics in current academia, has become a key technology to accurately analyze and forecast time-series data. Among all these computational intelligence methods, deep learning is in evidence [12]. Deep learning is a newly developed and fast-growing class of machine learning algorithms. A deep network has multiple hidden layers between the input and output layers in order to model complicated nonlinear relationships. With enough training materials, which usually are labeled data, the parameters in a deep network can be well trained to extract complex features from large data. Therefore, deep learning methods have been successfully applied in lots of fields, including scene understanding, natural language processing, self-driving, audio recognition, and so on. Because of its strong automatic feature extraction and pattern recognition ability, deep-learning-based methods are extremely suitable for electrical load forecasting, where lots of influences exist. The most similar approach to our proposed method is the short-term deep neural network (SDNN) model [13]. This model contains three steps: data preprocessing, network training, and forecasting. Historical weather conditions and load values are used as the network input. Compared to the aforementioned deep-learning-based approaches, our method adopts a specially designed two-step forecasting scheme, and takes into account various influences to analyze their impact.

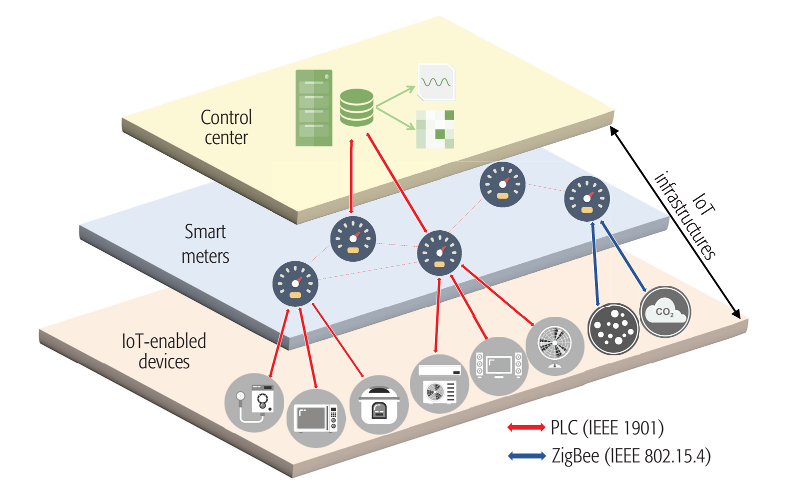


Figure 1. The load forecasting and analysis system based on IoT-enabled sensors and devices.

Forecasting System: Concept and Design

When starting with the research of electrical load, we wonder what on Earth are the possible influences, and which factors have a role in the load variation? We decide to begin with the analysis of the historical record and try to find some inspiration. Figure 2 presents an electrical load record of a large city in south China. The data is sampled every five minutes, from January 2014 to June 2016. As can be seen, there are some obvious patterns in the load variation. On one hand, these records reflect an annual periodicity. The power load peaks between June and October every year, and hits bottom around February, which has significant seasonal characteristics. On the other hand, there is also an obvious daily periodicity, that is, the load value stays high in the daytime and drastically drops at night.

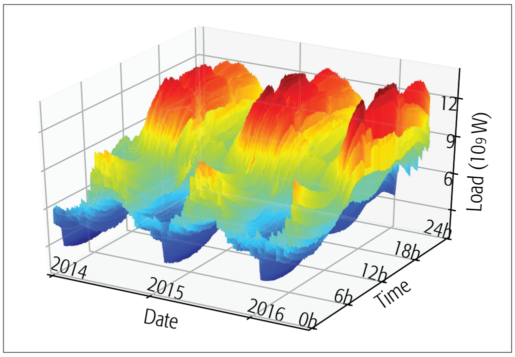


Figure 2. Urban electrical load in China (sampled every five minutes).

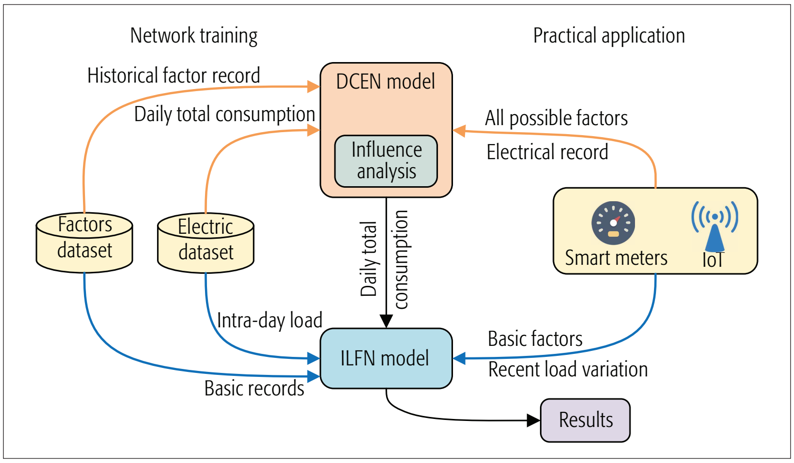


Figure 3. The framework of the proposed load forecasting system.

Although with some simple data analysis like this the presented load patterns can lead to a few preliminary conclusions, it is very difficult to truly understand the complex relationship between the power consumption and influence factors. In fact, weather and some other factors play much greater parts in electrical load variation, and also in more complicated ways, which are far beyond the capacity of humanity and traditional load forecasting methods. Besides, we empower smart meters with the ability to communicate with other IoT-enabled devices in the system, leading to more extensive input attributes. As shown in Fig. 1, the IoT infrastructure is a fundamental component, because it monitors the factors and sends the data to the control center. The IoT infrastructure consists of IoT-enabled devices, including smart meters, gadgets, and appliances, and the communication network. For economy and reliable communication, we adopt power line communication (PLC), which can transfer low-bit-rate data at low cost [14], and ZigBee, which can exchange data wirelessly within a 100 m range such as in a home or building, as the communication network. And smart meters, just like the sink nodes in a wireless sensor network (WSN), are responsible for collecting data from the household devices and sensors, and uploading the data to the control center every 30 s. The captured data is extremely complicated and contains lots of useful information. We desire to learn these patterns with a deep-learning-based system to give an accurate estimation of the future electrical load.

As mentioned above, we notice that several researchers have attempted to utilize deep learning in load forecasting, but many of them are facing a problem of low precision. Frequently, the existing approaches give inaccurate daily consumption readings even when they have the ability to predict short-term load precisely. This is a serious problem because many participants of electricity markets regard the *daily* consumption value as an important reference for decision making. A too large estimation may lead to energy waste, while a too small value can possibly cause an insufficient supply. We adopt a specially designed two-step forecasting scheme to address this problem.

The framework of the proposed load forecasting system is presented in Fig. 3. In our method, two individual models are used to predict the daily consumption and intra-day variation: the daily consumption estimation network (DCEN) and intra-day load forecasting network (ILFN). There are two major reasons that we design the two-step forecasting scheme. One reason is that the estimation value of DCEN is not only a helpful guide for electrical companies and consumers, but an important input to the ILFN model, which takes the daily consumption value as a reference and also a limitation. Therefore, with the proposed scheme, the estimated variation can be much closer to the actual load values. The other reason is that electrical load is influenced by various factors, which is usually in the unit of days, such as the daily maximum or minimum temperature, daily precipitation, daily sunshine duration, and, of course, the date. Also, the relevant data is most often obtained in the unit of days. Based on these facts, we concentrate all the possible factors at the DCEN model to accurately predict the daily consumption, and only adopt several basic factors to support the intra-day forecasting in order to simplify the network structure,

There are 10 hidden layers in the proposed DCEN model. Layers 1 and 2 have 4096 neurons each, layers 3~5 have 2048 neurons each, and layers 6~10 have 1024 neurons each. We envisaged implementing DCEN as an extremely complicated model to hold and analyze the super large data. However, we found that data engineering is a more efficient way for this task. With well selected input, even a common deep model can extract sufficient features and give meaningful information for the load forecasting. We demonstrate this in the experiment section.

As mentioned above, the key problem in the DCEN model is the selection and preprocessing of the input data. A massive amount of data is acquired by the proposed IoT system. Among them, we pick the following data as the input. As the most instructive reference, the daily consumption of the past seven days is selected; to reflect any periodic characteristics, the time relevant attributes are also adopted, including the date, Chinese lunar date, and day of the week; as the most important and complicated data, weather relevant attributes are of great significance to the DCEN model, including the temperature, air pressure, vapor pressure, precipitation, evaporation, wind speed, and sunshine duration. These data are preprocessed to give out the maximum, minimum, and average values, and then normalized to generate the final inputs, which include 7 electrical attributes, 3 time relevant attributes, and 22 weather relevant attributes.

After obtaining the daily consumption data, we adopt the ILFN model to estimate the intra-day load variations. ILFN is also a classic deep model with five hidden layers, and each layer has 512 neurons. The difference is that ILFN needs fewer input attributes compared to the DCEN model, because all the possible influences have been handled by DCEN, and most of them can be neglected in ILFN. The input only includes several basic factors and the recent load variation. In detail, the input data includes the estimated daily consumption value, the time relevant attributes, the load values in the last five time units, and some relevant readings. DCEN and ILFN are performed for each electrical consumption unit for more nuanced and accurate forecasting. This is mainly benefited by the lower-granularity data from IoT-enabled systems. Table 1 gives the comparison between traditional systems and the IoT-based system. It can be seen that the proposed system is able to monitor the detailed information of the residents' house, and give solid data support for the forecasting system. These valuable data can be a useful addition to the records captured by the onboard sensors, such as the operation log of smart appliances, which can be an important reference for the residents' energy usage habits. As one of the most important factors, some detailed weather condition data can only be captured by household sensors, such as the indoor temperature, sunshine duration, and indoor air quality, which differentiate in every house but have a strong effect on the energy consumption.

	Existing systems	IoT-enabled system
Data source	Onboard sensors	IoT devices
Granularity	Community	House/room
Data scope	Limited sensors	Extended by devices
Controllability	Controlled by provider	Controlled by residents
Deployment cost	Expensive sensors	Low-cost sharing
Adaptability	Fully applicable	Available in IoT network

Table 1. Features comparison of forecasting systems.

Influence Analysis

We need to go a step further than merely implementing a forecasting system. Although the proposed DCEN and ILFN model can make accurate predictions, it is no doubt necessary to figure out the mechanism behind the network structure, rather than simply leaving it as a black box. We start by clarifying the focus of the system, in other words, what the system is really concerned about among all the input attributes, including the historical load, weather factors, and time relevant information. This is very important not only for this research, but for other load forecasting applications in different areas, because the analysis of the forecasting mechanism can serve as useful guidance for system design. For example, the 32 attributes we select in the proposed instance are probably not suitable for other smart grids, especially the Chinese lunar date, which is only of significance to some areas in China. So how do we find the "right" attributes for a specific area? Influence analysis is an efficient way to perform this task. In the stage of system design, researchers can push all the possible factors into a prototype system, and after adequate training, analyze the contributions of each attribute. Ultimately, the researchers are able to obtain the accurate combination of possible factors. It is an economic solution to know the factors that truly matter before the large-scale deployment of smart meters and sensors. Besides, following the trend of IoT, an increasing number of devices will be IoT-enabled. Therefore, smart meters will get much more complicated input data in the future, and the influence analysis will play a key role in discriminating the value of various data sources. In addition, influence analysis can also be used as a technical measure for the network overfitting, which is a common and serious problem in the training process, but with few effective measurement means for a long time. Overfitting frequently occurs when the deep model is too complicated while having insufficient input data. An overfitted model may have good statistical results on the training materials, but usually perform poorly on actual applications, due to its overreaction to minor fluctuations. Through influence analysis, researchers can obtain some information regarding whether overfitting occurs or not. This is mainly because an overfitted network usually extracts meaningless features from the raw data, which are impossible to comprehend in most cases. On the contrary, well trained networks analyze the data in a human-like way. This difference can become an effective standard of distinguishing overfitted networks from normal ones.

For these purposes, we design a novel visualization method to analyze the contributions from each input attribute to the final output. We notice that a trained network has fixed parameters, including weights and biases. Therefore, the final output is merely related to the input. And if we change one input unit of an input attribute, the output result will also be changed, which provides a way to infer the contribution of one single input attribute. The analysis algorithm is explained below. First, each attribute in each input sample is fine-tuned to generate new output results. Then each new output value is compared to the former results to show their own contributions. At last, the normalized differences are presented in heatmap form.

An example of the proposed influence analysis is shown in Fig. 4. The analysis is conducted with a well trained network. For simplicity, only some relevant attributes that have significant influence on the final forecasting result are presented in the figure, including the date, the Chinese lunar calendar date, the day of the week, the temperature, and the air pressure. The influences are shown in color. The red areas have greater influence than the blue areas. Since all the attributes are normalized and change to the same extent, the generated heatmap can give an intuitive representation regarding which parts of the attributes have the most influence on the forecasting results.

Among all the presented heatmaps, the temperature attribute has the most significant effect on the final output, according to the highlighted zone around 25°C in the temperature channel. As can be seen, the highlighted zone is converged around 25°C, which is mainly because this is a sensitive cutoff point to determine whether to use the air conditioner. When the temperature is lower than 25°C, there is no cooling need. On the other hand, when the temperature is much higher than 25°C, the cooling need always exists, and a minor temperature change has little influence

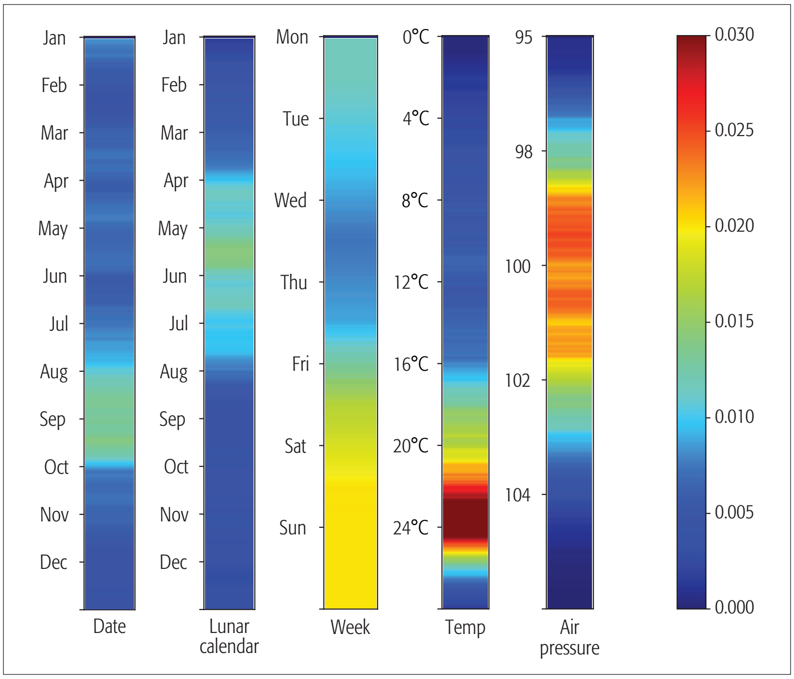


Figure 4. The generated heatmap for influence analysis.

on the power consumption. Traditionally, there is no demand for heating in south China. Therefore, low temperature also has little influence on the electrical load. We are very surprised at the rationality and interpretability when we see the visualization results for the first time. Not only the temperature but other attributes show meaningful heatmaps. For example, the date channel and lunar date channel both have highlighted zones around several legal holidays, when the factories are usually closed and, as a result, the electrical load drops. Lunar date is a traditional calendar in China, and many holidays are based on lunar date. Therefore, we set the attribute of lunar date to reflect some specific periodic patterns in China. In the week channel, the influence value of weekends is higher than that of weekdays, because the weekends are also rest days for many industries. As for the air pressure, according to much existing research, there is a strong inverse correlation between the air pressure value and electrical load. This is because the lower air pressure frequently results in oppressive weather, which fuels the increase in cooling needs. Besides the channels presented in the figure, we also analyze the influence of historical load data, that is, the daily consumption of the past seven days. Their normalized influence values are 0.07, 0.011, 0.009, 0.007, 0.005, 0.008, and 0.007, respectively, for the past days from yesterday to seven days ago. It can be seen that the closest point in time has the most significant effect on the forecasting result.

As expected, the visualization results demonstrate that the proposed system can draw rational conclusions with an intelligible inferential process. The analysis method enables researchers to select attribute combinations and judge overfitting networks.

Performance Evaluation

To show the actual forecasting performance of our method and demonstrate the effectiveness of the specially designed two-step forecasting scheme, several simulations are conducted in this section. The input is the actual record of an urban area in China. We create an instance [15] of the proposed models and train the system with the input data. As shown in Fig. 5a, we perform two forecasting tests in a period of one hour. The red line indicates the forecasting results generated with both the DCEN and ILFN models, the green line represents the results generated with only the ILFN model, and the dotted line is the actual load value. It can be seen that although both forecasting lines are close to the truth value, the green line shows some offset as a whole when the two-step forecasting is not adopted. More precisely, nearly all prediction values in the green line are bigger than the actual value, leading to inaccurate daily total consumption that is much bigger than the true value. In contrast, the red line is well distributed on both sides of the dotted line, which may result in more accurate total consumption. A quantitative analysis experiment is performed to give a precise performance comparison between these two tests, also with two other existing approaches: the state-of-the-art deep-learning-based SDNN model [13] working on the same 32 attributes including electrical data, time relevant data, and weather data, and the classical HW method merely working with an electrical record.

We adopt three mathematical indexes to quantitatively measure their performance. Figure 5b gives the comparison result. The mean absolute percentage error (MAPE) is a famous measure of forecasting precision in statistics. MAPE is scale-independent, and is favored when comparing prediction accuracy between different datasets. The root mean square deviation (RMSD) is another accuracy index, but can only be used to compare prediction errors of different models for the same dataset as it is scale-dependent. RMSD is normalized with the mean value of the measurements in this article, namely, coefficient of variation of the RMSD (cvRMSD). In addition, we design a new measure named total consumption relative error (TCRE) to show the effect of our two-step forecasting scheme. TCRE is calculated using the actual daily consumption value and the sum of all estimations in one day. All of these three measurements are expressed as percentages.

In Fig. 5b, the first group represents the result of the proposed two-step approach, the second group indicates the prediction without DCEN model, the third group is the SDNN model, and the last group represents the HW method. As can be seen, even without DCEN, the proposed method can achieve state-of-the-art performance similar to SDNN. However, it is significantly outperformed by the proposed two-step approach in the measure of TCRE. These numerical results once again demonstrate the necessity and effectiveness of our two-step forecasting method. Compared to other approaches, the proposed method performs much better in the prediction precision of both intra-day load variation and daily total consumption.

To demonstrate the effect of the extensive data from IoT-enabled devices, we perform an additional comparison experiment in a residential building in the same city. As shown in Fig. 5c, the first group is the results of the proposed method using lower-granularity data, which is monitored per house; while the second group represents the results using higher-granularity data, which is captured as the

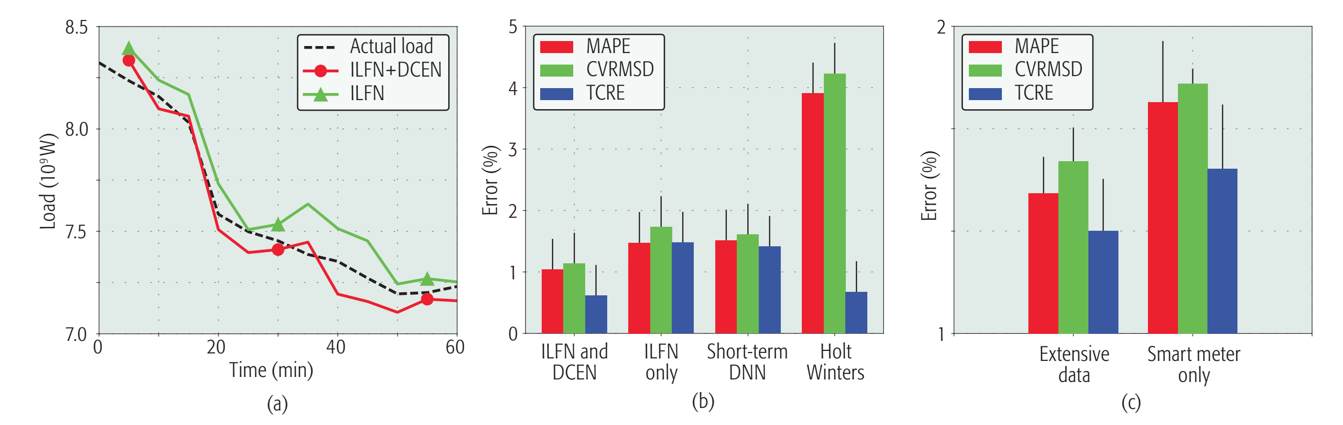


Figure 5. Evaluation results of the forecasting methods: a) forecasting results (12 predictions in 60 minutes); b) comparison of forecasting precision; c) comparison in an IoT-enabled building.

whole building. We can see the first group outperforms the second one in all of the three indexes. This improvement can be attributed to the differences of temperature, humidity, sunshine duration, and indoor air quality among the rooms in the building. Through IoT-enabled devices, the system obtains the ability to accurately forecast the energy consumption for every electrical unit, and as a result, improve its prediction precision of the total consumption.

Conclusion

An IoT-based electrical load forecasting method is proposed in the article. A huge advantage of the method is its two-step forecasting scheme, which significantly increases the prediction precision for daily total consumption. Another major difference is that we adopt deep learning methods to learn complicated patterns that form all the possible influences, and achieve a state-of-the-art forecasting performance in the evaluations. In addition, we also propose an analysis method to find the relationship between the influences and the electrical load, and design a heatmap generation method to show the specific impacts of each attribute on forecasting results. This analysis method is also of much guiding significance for smart grids in other countries, especially for ones with vast territory and varied climates. The results prove its effectiveness.

One limitation is that in the proposed system, a huge amount of data needs to be transferred on the communication network, which can bring a big challenge to the existing infrastructures. One feasible solution is to adopt edge servers near the client side for better computing balance and less communication cost, which is also included in our future works.

Acknowledgment

This work is partially supported by JSPS KAKENHI Grant Number JP16K00117, JP15K15976, and the KDDI Foundation. Mianxiong Dong is the corresponding author.

References

[1] L. Xiao *et al.*, "Research and Application of a Hybrid Model Based on Multi-Objective Optimization for Electrical Load Forecasting," *Applied Energy*, vol. 180, 2016, pp. 213–33.

[2] D. Alahakoon and X. Yu, "Smart Electricity Meter Data Intelligence for Future Energy Systems: A Survey," *IEEE Trans. Industrial Informatics*, vol. 12, no. 1, Feb 2016, pp. 425–36.

[3] H. Li *et al.*, "Mobile Crowdsensing in Software Defined Opportunistic Networks," *IEEE Commun. Mag.*, vol. 55, no. 6, 2017, pp. 140–45.

[4] K. S. Cetin and Z. O'Neill, "Smart Meters and Smart Devices in Buildings: A Review of Recent Progress and Influence on Electricity Use and Peak Demand," *Current Sustainable/Renewable Energy Reports*, vol. 4, no. 1, 2017, pp. 1–7.

[5] L. Hernandez *et al.*, "A Survey on Electric Power Demand Forecasting: Future Trends in Smart Grids, Microgrids and Smart Buildings," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 3, 3rd qtr. 2014, pp. 1460–95.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, 2015, pp. 436–44.

[7] J. Lloret *et al.*, "An Integrated Iot Architecture for Smart Metering," *IEEE Commun. Mag.*, vol. 54, no. 12, Dec. 2016, pp. 50–57.

[8] J. J. Nielsen *et al.*, "What Can Wireless Cellular Technologies Do About the Upcoming Smart Metering Traffic?," *IEEE Commun. Mag.*, vol. 53, no. 9, Sept. 2015, pp. 41–47.

[9] J. Jiang and Y. Qian, "Distributed Communication Architecture for Smart Grid Applications," *IEEE Commun. Mag.*, vol. 54, no. 12, Dec. 2016, pp. 60–67.

[10] H. Li *et al.*, "Multimedia Processing Pricing Strategy in GPU-Accelerated Cloud Computing," *IEEE Trans. Cloud Computing*, 2017. DOI: 10.1109/TCC.2017.2672554.

[11] M. Rana and I. Koprinska, "Forecasting Electricity Load with Advanced Wavelet Neural Networks," *Neurocomputing*, vol. 182, 2016, pp. 118–32.

[12] I. M. Coelho *et al.*, "A GPU Deep Learning Metaheuristic Based Model for Time Series Forecasting," *Applied Energy*, vol. 201, 2017, pp. 412–18.

[13] S. Ryu, J. Noh, and H. Kim, "Deep Neural Network Based Demand Side Short Term Load Forecasting," *2016 IEEE Int'l. Conf. Smart Grid Commun.*, Nov 2016, pp. 308–13.

[14] M. Yigit *et al.*, "Power Line Communication Technologies for Smart Grid Applications: A Review of Advances and Challenges," *Computer Networks*, vol. 70, 2014, pp. 366–83.

[15] Y. Jia, E. Shelhamer *et al.*, "Caffe: Convolutional Architecture for Fast Feature Embedding," *Proc. 22Nd ACM Int'l. Conf. Multimedia*, ser. MM '14, ACM, 2014, pp. 675–78.

Biographies

Liangzhi Li (16096502@mmm.muroran-it.ac.jp) received his B.Sc. and M.Eng. degrees in computer science from South China University of Technology (SCUT) in 2012 and 2016, respectively. He is currently pursuing a Ph.D. degree in electrical engineering at Muroran Institute of Technology, Japan. His main fields of research interest include machine learning, big data, and smart grid. He has received the best paper award from FCST 2017.

Kaoru Ota (ota@mmm.muroran-it.ac.jp) received her M.S. degree in computer science from Oklahoma State University in 2008, and B.S. and Ph.D. degrees in computer science and engineering from the University of Aizu, Japan, in 2006 and 2012, respectively. She is currently an assistant professor with the Department of Information and Electronic Engineering, Muroran Institute of Technology. She serves as an Editor for *IEEE Communications Letters*.

Mianxiong Dong (mxdong@mmm.muroran-it.ac.jp) received his B.S., M.S., and Ph.D. in computer science and engineering from the University of Aizu. He is currently an associate professor in the Department of Information and Electronic Engineering at Muroran Institute of Technology. He serves as an Editor for *IEEE Communications Surveys & Tutorials*, *IEEE Network*, *IEEE Wireless Communications Letters*, *IEEE Cloud Computing*, and *IEEE Access*.

Software Defined Machine-to-Machine Communication for Smart Energy Management

Zhenyu Zhou, Jie Gong, Yejun He, and Yan Zhang

The authors provide a comprehensive review of the state-of-the-art contributions from the perspective of SDN and M2M integration. The overall design of the proposed software-defined M2M (SD-M2M) framework is presented, with an emphasis on its technical contributions to cost reduction, fine granularity resource allocation, and end-to-end quality of service guarantee.

Abstract

The successful realization of smart energy management relies on ubiquitous and reliable information exchange among millions of sensors and actuators deployed in the field with little or no human intervention. This motivates us to propose a unified communication framework for smart energy management by exploring the integration of software-defined networking with machine-to-machine communication. In this article, first we provide a comprehensive review of the state-of-the-art contributions from the perspective of software defined networking and machine-to-machine integration. Second, the overall design of the proposed software-defined machine-to-machine (SD-M2M) framework is presented, with an emphasis on its technical contributions to cost reduction, fine granularity resource allocation, and end-to-end quality of service guarantee. Then a case study is conducted for an electric vehicle energy management system to validate the proposed SD-M2M framework. Finally, we identify several open issues and present key research opportunities.

Introduction

Despite the unprecedented development in the energy industry, the conventional energy system with centralized energy generation and unidirectional energy flows has become a bottleneck for facilitating the large-scale penetration of distributed and diversified renewable energy sources. Considering the intermittent and fluctuating characteristics of renewable energy and the stochastic charging/discharging behaviors and load profiles of electric vehicles, the high-level integration of uncontrolled and uncoordinated renewable generators and electric vehicles with the distribution networks will dramatically increase system volatility and disturbances, which often lead to power blackouts and brownouts due to cascading failures. Hence, smart energy management is urgently required to harness the huge potential of widespread renewable energy sources by dynamically optimizing the balance between energy supply and demand.

The successful realization of smart energy management lies in the real-time information of load-supply profiles and system operating conditions. By integrating every piece of the energy system with novel information and communication technologies (ICT), frequently updated measurements and samples of energy generation, transmission, distribution, storage, and consumption statuses can be retrieved via millions of sensors and actuators in the field. In particular, high-level syntactic and semantic interoperability among heterogeneous systems is enabled through open plug-and-play communication interfaces, which also provide the flexible control and self-deployment of standard and modular autonomous energy sources while hiding the diversity of underlying technologies. However, the communication network required by smart energy management is very different from conventional human-to-human-oriented telecommunication networks. This type of communication, which is characterized by ubiquitous information exchange among a large number of intelligent machines such as sensors, actuators, intelligent electronic devices (IEDs), smart meters, and so on, with little or no human intervention, is commonly known as machine-to-machine (M2M) communication [1].

Cellular technologies have become major driving forces for M2M due to the ubiquitous presence of cellular infrastructures and the availability of large-capacity long-range wide access. The Third Generation Partnership Project (3GPP) has specified several methods to support M2M communication in the current Long Term Evolution-Advanced (LTE-A) systems [2]. M2M devices are allowed to coexist with human-to-human communications in the same network and use random access channel (RACH) to build connections with centralized base stations (BSs). As a result, M2M communication, with self-organization, self-configuration, and self-healing capabilities, is expected to be a key enabler for the reliable operation of smart energy management.

Nevertheless, the seamless integration of M2M communication with smart energy management is still nontrivial. First of all, the conventional application-oriented method requires complete customization of M2M platforms for the specific application scenario, which has little flexibility in adapting to rapidly changing demand. It is extremely inefficient to manage the massive

This work was partially supported by the National Science Foundation of China (NSFC) under Grant Numbers 61601181 and 61372077; the Fundamental Research Funds for the Central Universities under Grant Number 2017MS13; the Beijing Natural Science Foundation (4174104); Beijing Outstanding Young Talent under Grant Number 2016000020124G081. This work is also partially supported by the projects 240079/F20 funded by the Research Council of Norway.

Digital Object Identifier: 10.1109/MCOM.2017.1700169

Zhenyu Zhou is with North China Electric Power University; Jie Gong is with Sun Yat-sen University; Yejun He is with Shenzhen University; Yan Zhang is with University of Oslo.

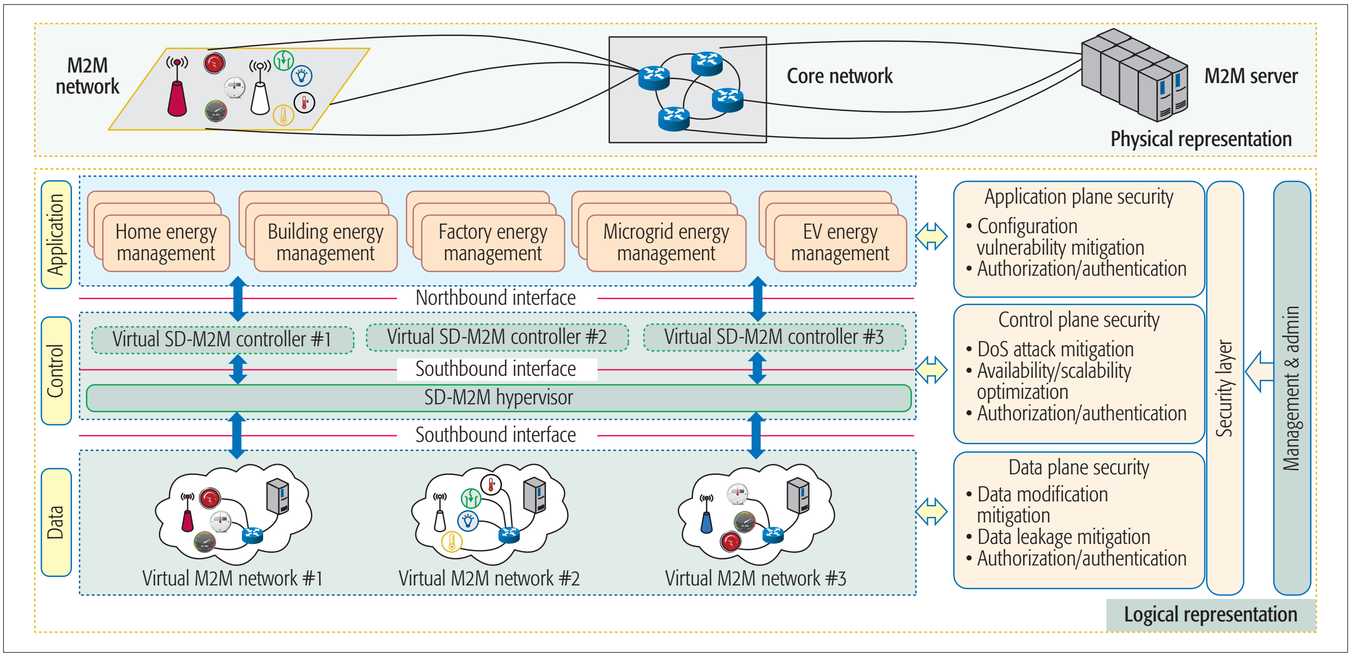

M2M network
Core network
M2M server
Physical representation
Application
Home energy management
Building energy management
Factory energy management
Microgrid energy management
EV energy management
Northbound interface
Control
Virtual SD-M2M controller #1
Virtual SD-M2M controller #2
Virtual SD-M2M controller #3
Southbound interface
SD-M2M hypervisor
Southbound interface
Data
Virtual M2M network #1
Virtual M2M network #2
Virtual M2M network #3
Application plane security
• Configuration vulnerability mitigation
• Authorization/authentication
Control plane security
• DoS attack mitigation
• Availability/scalability optimization
• Authorization/authentication
Data plane security
• Data modification mitigation
• Data leakage mitigation
• Authorization/authentication
Security layer
Management & admin
Logical representation

Figure 1. The conceptual architecture of software-defined M2M communication.

number of M2M devices in this way due to the increasing system complexity, and the extensive heterogeneity across hardware, interconnectivity, and deployment scenarios. Second, the tight coupling between applications and the task-oriented hardware provides little possibility of reusing existing physical M2M infrastructure for novel applications. Redundant hardware deployments are required for different applications or even the same application of different operators, which leads to excessively high capital and maintenance costs. Last but not least, power grid applications have diverse quality of service (QoS) requirements in terms of latency, burst size, throughput, and packet arrival rate. The coexistence of protection, control, monitoring, and billing traffic in the same communication network poses new challenges for efficient resource allocation design in M2M communication [3]. It is infeasible to realize intelligent resource allocation if the applied control logic is embedded into hardware devices.

Software-defined networking (SDN) provides an open architecture for enabling centralized control and automatic management of networks through the decoupling of the control plane and data plane, and the incorporation of network programmable capability. The design, deployment, management, and maintenance of networks can easily be implemented on an open-standard-based centralized controller rather than directly configuring a massive number of heterogeneous devices. There are some works attempting to integrate M2M with SDN. A virtual resource allocation algorithm was proposed for M2M communication underlaying software-defined cellular networks in [4]. Vukobratovic *et al.* presented a reconfigurable architecture for adapting Internet of Things (IoT) data transfer to subsequent data analysis based on the concept of network function computation [5]. Ameigeiras *et al.* designed an SDN-based M2M access cloud architecture to improve transmission latency, network scalability, and mobility support [6]. A software-defined dynamic M2M server selection and traffic redirection algorithm was proposed in [7] for a virtual home gateway. Hasegawa *et al.* proposed a joint bearer aggregation and control-data plane separation scheme to increase capacity of an M2M core network in [8]. There are some surveys [9, 10] that cover software-defined IoT at large. However, the above works mainly focus on conventional M2M networks. The specific technological characteristics and application scenarios when deploying SDN-based M2M for smart energy management have been largely neglected.

In this article, we present our visions on software-defined M2M (SD-M2M) communication, and study its potential in smart energy management. We start by introducing the overall design of the proposed SD-M2M architecture, with an emphasis on its technical contributions to intelligent service orchestration and resource allocation. Then the proposed SD-M2M framework is able to significantly reduce service, providing low cost, guaranteed end-to-end QoS delivery, and fine-granularity resource allocation by separating the data and control planes and decoupling service provision from physical infrastructures. We discuss how to integrate SD-M2M with different applications of smart energy management, and present a case study to evaluate the performance gains in both data delivery and energy management. Finally, we conclude the article and present major research open issues.

The Proposed Software-Defined M2M Framework

This section provides a detailed illustration of the proposed SD-M2M architecture, with a particular emphasis on its technical contributions to complexity reduction, fine-granularity resource allocation, and end-to-end QoS guarantee.

Figure 1 shows the SD-M2M architecture, which can be divided into four different planes: the data plane, the control plane, the application-

The centralized SD-M2M controller makes decisions on an up-to-date global view of the network state, and enables vendor-independent control over the corresponding virtual M2M network from a single logical point. This allows the implementation of fine-granularity control policies with enhanced network resource utilization efficiency and QoS provisioning capabilities.

tion plane, and the management and administration plane. The data plane is composed of all of the programable field equipment and network elements involved in M2M communication, including sensors, actuators, IEDs, smart meters, gateways, BSs, switches, routers, and so on. These are essential to support autonomous data acquisition and transmission in smart energy management. With the data-control decoupling, the data plane devices are greatly simplified without the need to understand hundreds of communication protocols.

The control plane consists of an SD-M2M hypervisor and multiple heterogeneous or homogeneous SD-M2M controllers. The virtualization of the physical M2M network is enabled by inserting a hypervisor between the data-plane devices and the controllers. The hypervisor views and interacts with the data-plane devices through the standard-based southbound interface and slices the abstracted physical infrastructures into multiple isolated virtual M2M networks that are controlled by their respective controllers. The hypervisor also sends the abstraction information to the controllers through the southbound interface. The centralized SD-M2M controller makes decisions on an up-to-date global view of the network state, and enables vendor-independent control over the corresponding virtual M2M network from a single logical point. This allows the implementation of fine-granularity control policies with enhanced network resource utilization efficiency and QoS provisioning capabilities.

The application plane covers an array of smart energy management applications such as home energy management (HEM), factory energy management (FEM), building energy management (BEM), microgrid energy management (MEM), and electric vehicle energy management (EVEM). With standard-based application programming interfaces (APIs) between the control and application planes, smart energy management applications can explicitly and programmatically communicate their requirements to the respective controllers via the northbound interface, and can thus operate on an abstraction of the M2M networks without being tied to the details of physical infrastructures.

The management and administration plane provides management and access control functions to all the other three planes (i.e., the data plane, the control plane, and the application plane). It covers static tasks including device setup and management, privacy and security policy configuration, firmware and software updates, performance monitoring, and so on. The security layer protects the data plane from various security threats such as flow rule modification, unauthorized access control, and side channel attack. In the control plane, the security layer provides solutions for controller access authorization and authentication, denial of service (DoS) or distributed DoS (DDoS) attack mitigation, controller availability and scalability optimization, and so on. Furthermore, security enforcement mechanisms can be implemented to secure the application plane from unauthorized and unauthenticated applications, fraudulent rule insertion, configuration vulnerabilities, and other application-specific security threats.

The main benefits of SD-M2M are summarized as follows.

Reduced complexity and accelerated innovation. In the SD-M2M communication framework, the underlying physical infrastructures are abstracted from smart energy management applications, and the complicated decision-making functions are left to the centralized controller. The controller is designed to hide the hardware details from service orchestration and provisioning, and to manage the data-plane devices automatically and intelligently via common APIs. We focus on how to realize seamless integration between SDN and M2M for smart energy management by exploring existing SDN controllers rather than trying to reinvent the wheel. As a result, SD-M2M provides unprecedented flexibility, programmability, and controllability for vendors and operators to build highly scalable and reliable M2M networks that can swiftly adapt to evolving applications of smart energy management. Rapid innovation is also enabled through the ability to tailor the behavior of the network and to deliver new applications and service differentiation in real time without the need to deploy and configure individual hardware devices.

End-to-end QoS guarantee in heterogeneous networks. Smart energy management functionalities such as real-time supervisory control and data acquisition, generation dispatch and control, and energy scheduling and accounting have diverse QoS requirements and different operation domains [11]. Thus, the most important challenge for conventional M2M communication is how to guarantee end-to-end QoS for different applications in heterogeneous networks. To provide a solution, SD-M2M creates a unified QoS delivery platform by decoupling the service provision functions from physical infrastructure domains. In this platform, the network resources and control functionalities are abstracted and sliced into distinct virtual networks, which are provided for respective applications via standard APIs. To guarantee reliable service delivery, the most appropriate virtual network is selected to meet the end-to-end QoS requirement. In this way, multiple virtual networks can be constructed on the same platform to meet the diverse QoS requirements of different system functions. The capability of inter-domain service delivery is significantly enhanced through the ability to coordinate network control and orchestrate resource allocation among controllers in different domains.

Fine-granularity resource allocation in a multitenant environment. In SD-M2M, physical infrastructures are abstracted from three dimensions of attributes: topology, physical device resources, and physical link resources. We focus on how to realize M2M infrastructure abstraction for intelligent service orchestration and resource allocation rather than redesign the concept of network functions virtualization. The degree of abstraction for each dimensional attribute can be flexibly controlled by the adjustment of physical resources. First, in the abstraction of network topology, the degree of abstraction relies on the virtual nodes and links. For instance, a physical topology that represents the layout of connected devices can be either abstracted as an identical virtual topology in the lowest degree of abstraction, or as a

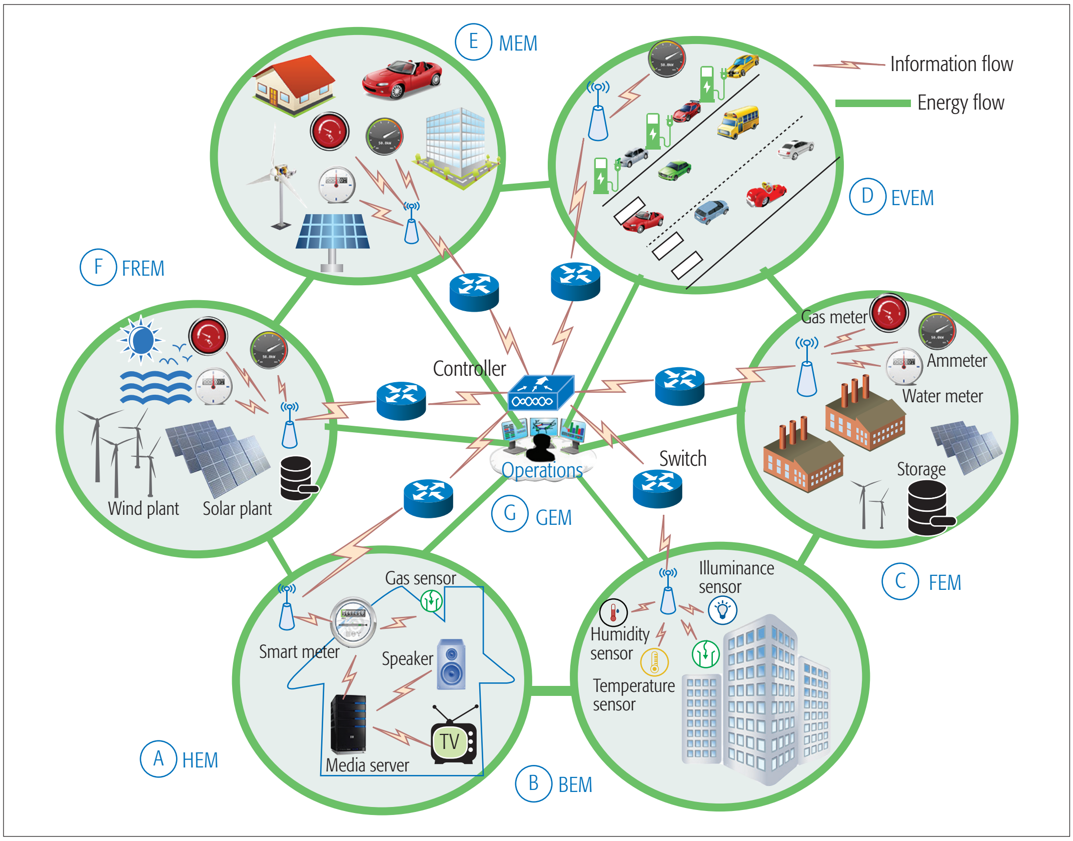


Figure 2. The scenario of deploying SD-M2M for smart energy management applications.

Smart BEM is realized by dynamically optimizing the energy consumption related to heating, cooling, ventilation, and lighting. SD-M2M provides a comprehensive platform to interact with M2M devices deployed in various building monitoring, control and automation systems, which are usually developed based on different communication protocols.

single virtual node or link in the highest degree of abstraction. Second, the degree of abstraction for physical device resources is dependent on CPU, memory, storage, and other computing resources. Third, in the abstraction of physical link resources, the ability to choose different levels of abstraction is determined by the allocation of link bandwidth, buffers, queues, and so on. Hence, SD-M2M offers a granular level of resource allocation in a highly abstracted and automated fashion, and allows the same physical infrastructures to be shared among multiple tenants.

Software-Defined M2M Communication for Smart Energy Management Applications

Figure 2 presents the scenario of deploying SD-M2M for smart energy management applications. We focus on several sub-areas where SD-M2M will play a key role and present how to integrate SD-M2M with different applications in a bottom-up approach. A comprehensive summary of the communication features and critical aspects for smart energy management applications is provided in Table 1.

Home Energy Management

HEM enables residential energy consumers to be actively involved in the grid operation through intelligent interaction with the external environment. Intelligent machines are embedded to collect home appliance operation status, energy consumption, home environment, and home user behaviors for smart HEM. SD-M2M will play a key role in facilitating HEM by shielding vendor-specific details and features of home appliances from application development and system operation. All of the registered M2M devices in a home can be divided into virtual networks with abstracted network, storage, and computing capability, and be managed through standard APIs to deliver home user demand-oriented services in a short time.

Building Energy Management

Residential and commercial buildings have been estimated to represent approximately half of the total world energy consumption. M2M communications are critical to collect real-time data of temperature, occupancy behavior, outdoor environment, humidity, illuminance, electricity price, and more. Smart BEM is realized by dynamically optimizing the energy consumption related to heating, cooling, ventilation, and lighting. SD-M2M provides a comprehensive platform to interact with M2M devices deployed in various building monitoring, control, and automation systems, which are usually developed based on different communication protocols. M2M networks in different buildings and systems can be abstracted and integrated into the same virtual network, which provides the benefit of allowing multiple buildings to be remotely managed by a centralized controller.

Factory Energy Management

Smart FEM will be a key enabler for the upcoming fourth industrial revolution. M2M devices are installed in a factory not only to collect energy generation, storage, and consumption data, but also to monitor real-time status of manufacturing lines. SD-M2M enables FEM operators to build

Application	Communication features	Critical aspects	Benefits of SD-M2M
Home energy management	• Delay-tolerant • Periodic/event-based • Short range • Low-level priority	• Diverse communication protocols • Massive connection • High random access loads • Small burst traffic	• Reduced cost and complexity • Accelerated innovation • Vendor-independent control
Building energy management	• Delay-tolerant • Periodic/event-based • Short range • Low-level priority	• Diverse communication protocols • Massive connection • Small burst traffic	• Reduced cost and complexity • Accelerated innovation • Coordinated management • Vendor-independent control
Factory energy management	• Delay-sensitive • Periodic/event-based • Middle-level priority • Middle range	• Diverse communication protocols • High reliability • Middle-level QoS requirement	• Reduced cost and complexity • Accelerated innovation • Coordinated management • Vendor-independent control
EV energy management	• Delay-sensitive • Semi-periodic/event-based • Middle-level priority • Middle range	• Mobility management • Random charging/discharing behaviors • High reliability • Middle-level QoS requirement	• Reduced cost and complexity • Coordinated mobility management • Fine-granularity resource allocation
Microgrid energy management	• Delay-sensitive • Semi-periodic/event-based • High-level priority • Middle range	• High reliability • High-level QoS requirement • Multi-tenant environment • Massive connection	• End-to-end QoS guarantee • Fine-granularity resource allocation • Coordinated management
Field renewable energy management	• Delay-sensitive • Periodic/event-based • Long range	• Fault-tolerant capability • High reliability • High-level QoS requirement	• End-to-end QoS guarantee • Low maintenance cost • Fine-granularity resource allocation
Grid energy management	• Extremely delay-sensitive • Periodic/event-based • No/limited retransmission • Long range	• Mission-critical • High reliability • Continuous transmission • High-level QoS requirement	• End-to-end QoS guarantee • Fine-granularity resource allocation • Coordinated management

Table 1. A comprehensive summary of the communication features and critical aspects for smart energy management applications.

highly reliable and programmable communication networks for integrating distributed renewable energy sources and energy-saving equipment such as motors and inverters. With the decoupling of data and control planes, the coexistence of energy and manufacturing traffic in the same communication network is supported through physical infrastructure abstraction and centralized resource coordination. Energy consumption improvement points and deterioration factors can easily be identified by interrelating product information with energy information through open and programmable APIs.

Electric Vehicle Energy Management

The massive amounts of data in every aspect of electric vehicles including locations, travel patterns, driver behaviors, battery states, and historical profiles are routinely collected for realizing smart EVEM, which reduces the energy demand-supply imbalance by absorbing excess energy during off-peak hours and discharging the batteries into the grid when needed [12]. SD-M2M with centralized intelligence provides an flexible communication network for coordinated charging and discharging different types of electric vehicles at distributed locations such as residential community, workplaces, parking lots, and charging stations. For mobility management, seamless handover of electric vehicles from one BS to another can be realized by coordinating network control and orchestrating resource allocation among multiple controllers.

Microgrid Energy Management

A microgrid is a small-scale electric power system with co-located distributed energy sources and loads. It can either synchronize to the main grid and operate in grid-connected mode, or operate in island mode by disconnecting both loads and energy sources from the main grid [13]. Hence, MEM provides the benefits of relieving the stress of load-supply imbalance through local consumption of distributed renewable energy sources. In SD-M2M, the physical M2M infrastructure can be abstracted and sliced into distinct virtual networks to support a variety of microgrid energy management functions with diverse communication requirements. Sufficient communication and computing resources should be allocated for the monitoring and control of critical interconnection points (i.e., the points of load connection, common coupling, and distributed energy source connection) in order to support seamless dispatch, scheduling, and control of distributed energy sources. Various stakeholders such as microgrid operator, distributed energy source operator and aggregator, and load aggregator are allowed to exchange key operating parameters in real time with the supported coordination among heterogeneous controllers.

Remote Field Renewable Energy Management

Large-scale solar and wind plants are normally deployed in remote and isolated renewable-rich areas such as deserts and offshore. M2M communication, which enables the reliable acquisition of field monitoring data over a long transmission range, serves as the basis for smart FREM. The data of temperature, pressure, humidity, solar position, and wind speed, as well as power quality-related parameters are collected and transmitted back to a control center for power output forecasting and energy management optimization. In remote harsh and hazardous locations, SD-M2M with fault-tolerant capability becomes an ideal choice. The decoupling of services and physical infrastructure makes it possible to reuse existing redundant devices during system malfunction. Furthermore, the centralized controller with a global view can easily detect device breakdowns and network disconnections to guarantee automatic acquisition of data with minimum interruptions.

Grid Energy Management

M2M devices such as phase measurement units are embedded into generation and transmission domain equipment to continuously collect critical data of power grid such as voltage, current, harmonics, and frequency [14]. These data are utilized by smart GEM to improve the flexibility and reliability performance of the overall power system. SD-M2M can support the real-time delivery of the strictly delay-sensitive system state measurements and high-resolution phase information. For instance, the integrated fiber-wireless communication infrastructures based on low-latency Ethernet passive optical networks and highly scalable cellular networks can be abstracted into different slices and then allocated at a fine-grained level to meet the end-to-end QoS requirements.

Case Study and Analysis

To validate the benefits of SD-M2M, we consider the application scenario of EVEM as shown in Fig. 2, which is composed of one gas generator, four wind turbines, and 100 electric vehicles. Important data such as charging time, location, battery state, and load profile are monitored by SD-M2M devices and sent back to M2M servers through cellular links. The case study is divided into two parts. In the first part, we evaluate the capability of SD-M2M for supporting real-time delivery of strictly delay-sensitive data. In the second part, we demonstrate the relationship between SD-M2M penetration rate and performance gain of smart energy management.

When an M2M device attempts to connect to a BS, it has to randomly select a preamble and send it to the BS via a time-frequency resource block. The BS decodes the received preamble and sends back a response message. A random access collision occurs if two or more M2M devices happen to select the same resource block, and then each M2M device has to wait for a random period and repeat random access again. Thus, the operation of smart energy management is in danger since critical data cannot be delivered immediately without delay. In particular, the probability of collision increases dramatically when a massive number of M2M devices attempt to access the network simultaneously.

SD-M2M provides a promising solution to the above challenge through an advanced level of resource abstraction and fine-granularity resource allocation. The hypervisor slices the physical infrastructure into K distinct virtual M2M networks based on QoS requirements. Without loss of generality, we focus on the kth ($k = 1, ..., K$) virtual network with N_k M2M devices. Assuming that M_k resource blocks are allocated by the controller, the total number of resource blocks is calculated as $\sum_{k=1}^{K} \mathrm{M}_k$. Given $K = 20$ and $M_k = 10$ for $k = 1, 2, ..., K$, the total number of required resource blocks is 200. Each M2M device only needs to be aware of the resource blocks allocated to the corresponding virtual network instead of sensing the whole physical network. If the achieved spectrum efficiency cannot meet the specified QoS requirement, more resources can be allocated to this virtual network for improving performance by coordinating resource allocation with other virtual networks. The study of inter-virtual network coordination is left for future study.

The strategy of each M2M device is to decide when to access the network and which resource block to choose. Since random access will be successful if and only if the resource block is idle and is not requested by others, the achievable spectrum efficiency is jointly determined by the number of available resource blocks, the actions of other M2M devices, and the channel quality of the requested resource block. As a result, each M2M device needs to decide whether or not to access the network at each time slot based on the state of resource blocks. A Markov decision process (MDP) provides an effective mathematical framework to formulate this category of decision making problems with a stochastic process. A standard MDP formulation involves the following elements: state, action, cost function, and state transmission. The system state S is defined as the set of all resource block states. The state transition probability can be modeled as a Poission process. The action is defined as the probability to access the network, which is relative to the system state. The optimization objective is to maximize the average transmission rate per device over the infinite time horizon. The MDP problem can be broken down into a collection of simpler subproblems and solved one by one via dynamic programming [15]. The proposed algorithm is guaranteed to obtain the optimal performance upon termination. The relative proof can be found in [15, references therein].

We compare the proposed algorithm with a baseline greedy algorithm in which each device always requests the resource block with the best channel quality. The results are shown in Figs. 3a and 3b. We consider a virtual network with $N_k = 100$ M2M devices. Figure 3a shows the average transmission rate per device with different numbers of resource blocks M_k. The proposed algorithm outperforms the greedy algorithm by more than 300 percent when $M_k = 10$. The reason is that the reuse gain of resource blocks is fully exploited. In Fig. 3b, we fix the total number of resource blocks $M_k = 6$, and change the maximum probability of accessing the network from 12 to 20 percent. It is shown that with the increase of maximum access probability, the performance degrades dramatically. The reason is that the collision probability increases exponentially as more devices attempt to access the network simultaneously. Nevertheless, the proposed algorithm still outperforms the greedy algorithm under all scenarios.

The decoupling of services and physical infrastructure makes it possible to reuse existing redundant devices during system malfunction. Furthermore, the centralized controller with a global view can easily detect device breakdowns and network disconnections to guarantee automatic acquisition of data with minimum interruptions.

SD-M2M provides great potential to achieve energy-efficient resource allocation through advanced level of physical resource abstraction and centralized control. However, such a benefit has yet to be fully harnessed due to the tradeoff between energy efficiency and other performance benchmarks such as spectrum efficiency and transmission latency.

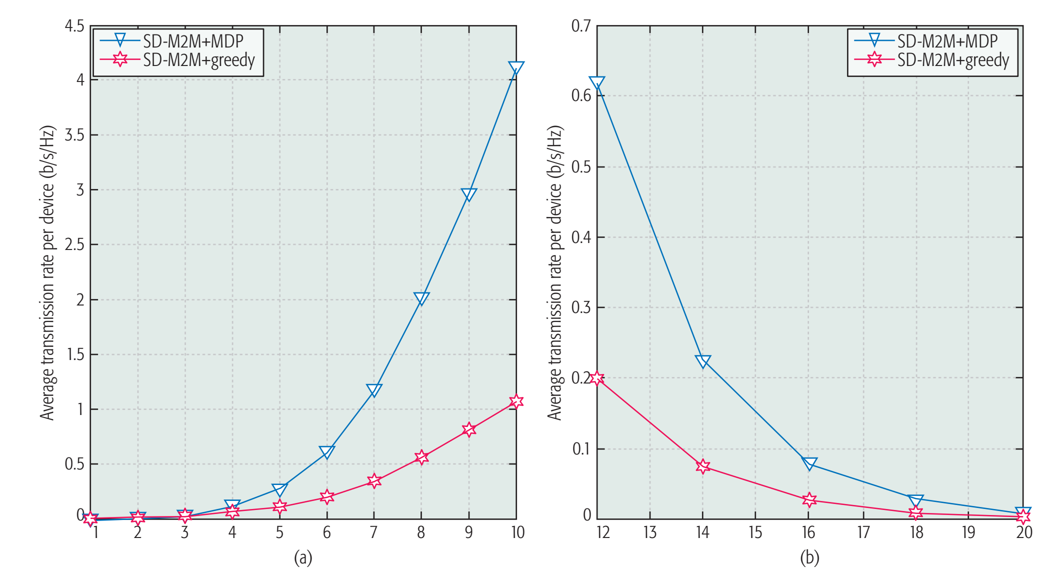


Figure 3. Spectral efficiency performance: a) average transmission rate per device vs. the number of total RBs; b) average transmission rate per device vs. the probability of accessing the network.

To evaluate the smart energy management performance, a robust energy scheduling approach proposed in our previous work [12] is employed. Robust energy scheduling allows a distribution-free model of uncertainties, and can efficiently alleviate the negative effect of data uncertainty. The goal is to minimize the generation cost of the gas generator under the constraints of active power balance, active power generation limits, charging and discharging power boundaries, charging demand balance, and spinning reserve. The optimization variables are when to charge and discharge electric vehicles, and the energy output of the gas generator. More details of the robust energy scheduling solution can be found in [12, references therein]. An electric vehicle cannot be scheduled if the critical data are not delivered on time, which occurs when either SD-M2M devices are not deployed or a QoS requirement is violated due to collision. We define the SD-M2M penetration rate as the ratio of electric vehicles that can be scheduled to the total number of electric vehicles.

Figure 4a shows the energy supply and demand profiles of electric vehicles, wind turbines, and local residents for a duration of 24 minutes. It is noted that the peak load starts at the sixth minute when the wind power output is low and the charging demand of electric vehicles is high. Figure 4b shows the total energy generation cost vs. the SD-M2M penetration rate. It is obvious that there is a positive correlation between cost reduction and penetration rate. For instance, the cost is reduced by 65 percent when the SD-M2M penetration rate is increased from 20 to 100 percent. It is interesting to note that the increment of SD-M2M penetration rate converts the exponential growth pattern of cost into a linear grown pattern. Based on the delay-sensitive mission-critical data delivered by SD-M2M, the peak load can be efficiently shifted by charging electric vehicles to absorb renewable energy during off-peak hours and discharging to produce energy during peak hours.

Conclusion and Open Issues

In this article, we propose a new software-defined M2M framework for emerging smart energy management applications. We review the current research progress on integrating SDN with M2M. Then the design principle of the proposed SD-M2M architecture is presented, and the technical contributions to cost and complexity reduction, end-to-end QoS guarantee, and fine-granularity resource allocation are elaborated in details. We also classify smart energy management applications into several classes based on operation domains, and provide a detailed treatment on how to integrate SD-M2M with each class of application. A case study is shown in an electric vehicle network to demonstrate the performance gains brought by SD-M2M in both spectral efficiency and energy management.

In the following, we point out four key research issues that call for more attention and efforts in the context of integrating SD-M2M with smart energy management.

Energy Efficiency and Energy Harvesting. Energy efficiency and energy harvesting are very important aspects of SD-M2M design due to limited battery capacity and high maintenance cost. SD-M2M provides great potential to achieve energy-efficient resource allocation through an advanced level of physical resource abstraction and centralized control. However, such a benefit has yet to be fully harnessed due to the tradeoff between energy efficiency and other performance benchmarks such as spectrum efficiency and transmission latency.

Dynamic Resource Virtualization and Sharing. The sharing of the same M2M infrastructures by multiple smart energy management operators calls for efficient resource slicing, isolation, and mapping algorithms. Due to the large number of optimization stages and scales, it is usually intractable to derive a polynomial-time solution for big instances of the formulated problem. Therefore,

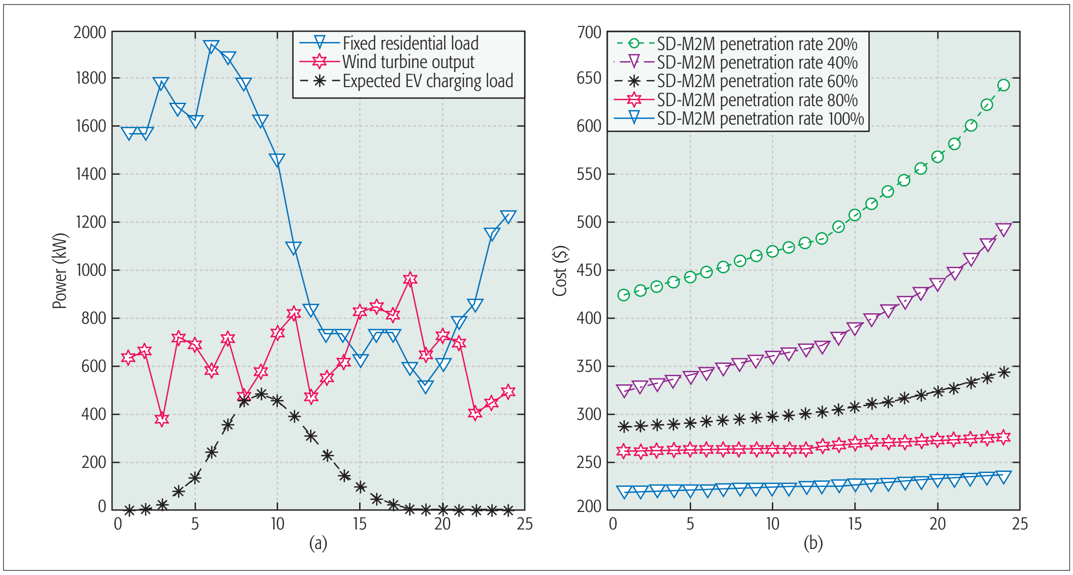


Figure 4. The smart energy management performance: a) the energy supply and demand profiles; b) the relationship between the total energy generation cost and the SD-M2M penetration rate.

The centralized SD-M2M controller is the single point of failure performance bottleneck, the promise of which leads to the collapse of both communication and energy networks. In particular, how to provide an efficient and seamless approach for privacy and trust management across a massive number of M2M devices is a valuable yet challenging issue.

alternative sub-optimal heuristic solutions should be investigated to tackle this challenge, and the corresponding optimality gap and computation complexity need to be analyzed in depth. Furthermore, considering the diverse or even conflicting objective functions of operators in multiple domains, game-theoretical or matching approaches should be incorporated to address the resource allocation problem.

Timescale Difference between Wireless Resource Allocation and Smart Energy Management. SD-M2M-based smart energy management confronts critical challenges caused by two-dimensional dynamics with different timescales. On one hand, wireless resource allocation is optimized according to dynamic channel variations on a timescale of milliseconds. On the other hand, energy utilization in smart grid is optimized based on dynamic load-supply profiles and electrical prices, which often vary on a timescale of hours, minutes, or seconds. Hence, there is lacking an efficient modeling approach to characterize the impacts of wireless resource allocation on smart energy management. The joint optimization of the two problems, which are on different timescales, requires further investigation.

Security Issues. The centralized SD-M2M controller is the single-point-of-failure performance bottleneck, the result of which leads to the collapse of both communication and energy networks. In particular, how to provide an efficient and seamless approach for privacy and trust management across a massive number of M2M devices is a valuable but challenging issue.

Acknowledgment

This work was supported in part by the National Science Foundation of China under Grant 61601181, in part by the Fundamental Research Funds for the Central Universities under Grant 2017MS13, in part by the Beijing Natural Science Foundation under Grant 4174104, in part by the Beijing Outstanding Young Talent under Grant Number 2016000020124G081, and in part by the projects 240079/F20 funded by the Research Council of Norway. This work was also supported in part by the National Natural Science Foundation of China under Grant 61372077, in part by the Science and Technology Innovation Commission of Shenzhen under Grant ZDSYS 201507031550105, and in part by Guangdong Provincial Science and Technology Programs under Grant 2013B090200011 and Grant 2016B090918080.

References

[1] Y. Zhang *et al.*, "Cognitive Machine-to-Machine Communications: Visions and Potentials for the Smart Grid," *IEEE Network*, vol. 26, no. 3, May 2012, pp. 6–13.

[2] A. Rico-Alvarino *et al.*, "An Overview of 3GPP Enhancements on Machine-to-Machine Communications," *IEEE Commun. Mag.*, vol. 54, no. 6, June 2016, pp. 14–21.

[3] Y. Yan et *al.*, "A Survey on Smart Grid Communication Infrastructures: Motivations, Requirements and Challenges," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 1, Feb. 2012, pp. 5-20.

[4] M. Li *et al.*, "Random Access and Virtual Resource Allocation in Software-Defined Cellular Networks with Machine-to-Machine (M2M) Communications," *IEEE Trans. Vehic. Tech.*, Dec. 2016, pp. 1–15.

[5] D. Vukobratovic *et al.*, " Condense: A Reconfigurable Knowledge Acquisition Architecture for Future 5G IoT," *IEEE Access*, vol. 4, July 2016, pp. 3360–78.

[6] P. Ameigeiras *et al.*, " Link-Level Access Cloud Architecture Design Based on SDN for 5G Networks," *IEEE Network*, vol. 29, no. 2, Mar. 2015, pp. 24–31.

[7] A. Papageorgiou *et al.*, "Dynamic M2M Device Attachment and Redirection in Virtual Home Gateway Environments," *IEEE ICC 2016*, Kuala Lumpur, Malaysia, July 2016, pp. 1–6.

[8] G. Hasegawa and M. Murata, "Joint Bearer Aggregation and Control-Data Plane Separation in LTE EPC for Increasing M2M Communication Capacity," *IEEE GLOBECOM 2015*, San Diego, CA, Dec. 2015, pp. 1–6.

[9] K. Sood, S. Yu, and Y. Xiang, "Software-Defined Wireless Networking Opportunities and Challenges for Internet-of-Things: A Review," *IEEE Internet of Things J.*, vol. 3, no. 4, Sept. 2016, pp. 453–63.

[10] I. Khan *et al.*, "Wireless Sensor Network Virtualization: A Survey," *IEEE Commun. Surveys and Tutorials*, vol. 18, no. 1, Mar. 2015, pp. 553–76.

[11] R. Yu *et al.*, "QoS Differential Scheduling in Cognitive-Radio-Based Smart Grid Networks: An Adaptive Dynamic Programming Approach," *IEEE Trans. Neural Network Learning Systems*, vol. 27, no. 2, Apr. 2015, pp. 435–43.

[12] Z. Zhou *et al.*, "Robust Energy Scheduling in Vehicle-to-Grid Networks," *IEEE Network*, vol. 31, no. 2, Mar. 2017, pp. 30–37.

[13] Z. Zhou *et al.*, "Game-Theoretical Energy Management for Energy Internet with Big Data-Based Renewable Power Forecasting," *IEEE Access*, Feb. 2017, pp. 1-14.

[14] G. C. Madueño *et al.*, "Assessment of LTE Wireless Access for Monitoring of Energy Distribution in the Smart Grid," *IEEE JSAC*, vol. 34, no. 3, Feb. 2016, pp. 675–88.

[15] J. Gong *et al.*, "Policy Optimization for Content Push Via Energy Harvesting Small Cells in Heterogeneous Networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 2, Nov. 2016, pp. 717–29.

Biographies

Zhenyu Zhou [M'11, SM'17] (zhenyu_zhou@ncepu.edu.cn) received his M.E. and Ph.D. degrees from Waseda University, Tokyo, Japan, in 2008 and 2011, respectively. Since March 2013, he has been an associate professor at North China Electric Power University. He received the Beijing Outstanding Young Talent in 2016 and the IET Premium Award in 2017. He is an Editor of *IEEE Access* and *IEEE Communications Magazine*. His research interests include green communications and smart grid.

Jie Gong [S'09, M'13] (gongj26@mail.sysu.edu.cn) received his B.S. and Ph.D. degrees in the Department of Electronic Engineering of Tsinghua University in 2008 and 2013, respectively. He visited the University of Edinburgh in 2012. During 2013–2015, he worked as a postdoctorial scholar at Tsinghua University. He is currently an associate research fellow in the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. His research interests include cloud RAN, energy harvesting technology, and green wireless communications.

Yejun He (heyejun@126.com) received his Ph.D.degree in information and communication engineering from Huazhong University of Science and Technology in 2005. He is a full professor with the College of Information Engineering, Shenzhen University, China, where he is the director of the Guangdong Engineering Research Center of Base Station Antennas and Propagation, and the director of the Shenzhen Key Laboratory of Antennas and Propagation. His research interests include wireless mobile communication, antennas, and radio frequency. He is a fellow of IET.

Yan Zhang [M'05, SM'10] (yanzhang@ieee.org) is a full professor at the Department of Informatics, University of Oslo, Norway. He is an Editor of *IEEE Communications Magazine*, *IEEE Transactions on Green Communications and Networking*, *IEEE Communications Surveys & Tutorials*, and others. His current research interests include next-generation wireless networks leading to 5G, green, and secure cyber-physical systems (e.g., smart grid, healthcare, and transport). He is an IEEE VTS Distinguished Lecturer. He is a Fellow of IET.

5G Mobile Cellular Networks: Enabling Distributed State Estimation for Smart Grids

Mirsad Cosovic, Achilleas Tsitsimelis, Dejan Vukobratovic, Javier Matamoros, and Carles Antón-Haro

The authors show how the emerging 5G mobile cellular network, with its evolution of machine-type communications and the concept of mobile edge computing, provides an adequate environment for distributed monitoring and control tasks in smart grids. In particular, they present in detail how smart grids could benefit from advanced distributed state estimation methods placed within the 5G environment.

Abstract

With the transition toward 5G, mobile cellular networks are evolving into a powerful platform for ubiquitous large-scale information acquisition, communication, storage, and processing. 5G will provide suitable services for mission-critical and real-time applications such as the ones envisioned in future smart grids. In this work, we show how the emerging 5G mobile cellular network, with its evolution of machine-type communications and the concept of mobile edge computing, provides an adequate environment for distributed monitoring and control tasks in smart grids. In particular, we present in detail how smart grids could benefit from advanced distributed state estimation methods placed within the 5G environment. We present an overview of emerging distributed state estimation solutions, focusing on those based on distributed optimization and probabilistic graphical models, and investigate their integration as part of the future 5G smart grid services.

Introduction

In recent years, two main trends have emerged in the evolution of power grids:

- The de-regulation of energy markets
- The increasing penetration of renewable energy sources

The former results in an increased exchange of large amounts of power between adjacent areas, possibly under the control of different regional utilities. The latter leads to larger system dynamics, due to the intermittency of renewable energy sources. To ensure power grid stability, such variations must be timely and accurately monitored.

State estimation (SE) is a key functionality of the electric power grid's energy management systems. SE aims to provide an estimate of the system state variables (voltage magnitude and angles) at *all* the buses of the electrical network from a set of remotely acquired measurements. The centralized (classical) SE schemes may prove inapplicable to emerging decentralized and dynamic power grids, due to large communication delays and high computational complexity that compromise their ability for real-time operation. Hence, the interest of the community is shifting from centralized to distributed SE algorithms based on more sophisticated optimization techniques beyond the classical iterative Gauss-Newton approaches [1].

Instrumental to this evolution is the deployment of synchronized phasor measurement units (PMUs) able to accurately measure voltage and current phasors at high sampling rates. Exploiting PMU inputs for a robust, decentralized, and real-time SE solution calls for novel communication infrastructure that would support the future wide area monitoring system (WAMS). WAMS aims to detect and counteract power grid disturbances in *real time*, thus requiring a communication infrastructure able to:

- Integrate PMU devices with extreme reliability and ultra-low (millisecond) latency
- Provide support for distributed and real-time computation architecture for future SE algorithms
- Provide backward compatibility to legacy measurements traditionally collected by supervisory control and data acquisition (SCADA) systems

The advent of fifth generation (5G) communication networks will largely facilitate the provision of the distributed information acquisition and processing services required in WAMS systems. As far as information *acquisition* is concerned, the introduction of massive machine-type communication (mMTC) services will allow for a large-scale deployment of advanced metering infrastructure (AMI). For those measurement devices (e.g., PMUs) requiring both very low latency and very high reliability, resorting to ultra-reliable low-latency communications (URLLC) services [2] will be needed. As for information *processing*, novel architectural concepts such as mobile edge computing (MEC) will be key for the deployment of the aforementioned distributed SE approaches [3].

The purpose of this work is twofold:

- To discuss the fundamental role to be played by 5G networks as an enabler of advanced distributed SE schemes
- To place two promising distributed SE solutions (based on distributed optimization and probabilistic graphical models) in such a 5G communications scenario

Specifically, first we describe how distributed SE can be integrated into the framework of MEC, while acquiring measurements via 5G MTC services. Then we focus on two distributed SE approaches based either on the alternating direction method of multipliers (ADMM) or on probabilistic graphical models and belief propagation

Digital Object Identifier:
10.1109/MCOM.2017.1700155

Mirsad Cosovic is with Schneider Electric DMS NS; Achilleas Tsitsimelis, Javier Matamoros, and Carles Antón-Haro are with Centre Tecnològic de Telecomunicacions de Catalunya (CTTC/iCERCA); Dejan Vukobratovic is with the University of Novi Sad.

0163-6804/17/$25.00 © 2017 IEEE

(BP) algorithms. We also assess their performance and discuss their applicability in realistic 5G communication scenarios, using the corresponding *centralized* versions of the SE schemes as benchmarks.

5G Enhancements for Distributed Information Acquisition and Processing in Smart Grids

In this section, we review 5G radio interface enhancements and MEC concepts as the main enablers for future 5G smart grid applications.

Radio Interface Enhancements for 5G

Third Generation Partnership Project (3GPP) standards providing radio interface enhancements for MTC have been recently adopted within 3GPP LTE Release 13. Three solutions for MTC services are introduced: enhanced MTC (eMTC), narrowband Internet of Things (NB-IoT), and extended coverage GSM Internet of Things (EC-GSM-IoT) [4]. For legacy (SCADA) measurement devices such as remote terminal units (RTUs), a suitable solution is provided by eMTC albeit with the same reliability and latency guarantees as provided by the LTE physical (PHY) layer. In contrast, for massive access of low-rate low-cost devices, new NB-IoT or EC-GSM-IoT extensions provide a proper solution. For example, NB-IoT targets up to 50,000 devices per macrocell with extended coverage, thus providing an ideal solution for smart meter data acquisition. However, significant improvements over 4G radio interface in both reliability and latency are needed for real-time services relying on PMUs.

3GPP 5G standardization of the New Radio (NR) interface is initiated in Release 13 with a requirement and architecture study. Following International Telecommunication Union (ITU) 5G requirements, NR will support two new services suitable for machine-type devices: mMTC and URLLC. mMTC will further enhance massive connectivity provisioning established by NB-IoT, targeting connection density of 10^6 and devices per square kilometer in an dense urban scenario while offering packet loss rates (PLRs) below 1 percent. For URLLC, the generic radio interface latency target is 0.5 ms, while reliability targets PLR of 10^{-5} for 32-byte packets and 1 ms latency. As we detail later, URLLC represents a suitable solution for WAMS real-time services that target system monitoring and control at the PMU sampling rates. However, in order to meet the stringent URLLC requirements, not only the radio interface but also the core network architecture will require novel solutions.

Mobile Edge Computing in 5G

The centralization and virtualization of core network functions within the so-called cloud RAN (C-RAN) architecture reduces costs and complexity of radio access network (RAN) densification. Further, the adoption of mobile cloud computing (MCC) architectures, allowing user equipment to offload computation and store data in remote cloud servers, facilitates the deployment of a number of novel user application and services. However, the major drawback associated with the MCC architecture is the large latency between the end user and the remote cloud center, thus limiting the applicability of MCC for services requiring very low latencies. This has led to a recent surge of interest in MEC architectures, where cloud computing and storage is distributed and pushed toward the mobile network edge [3]. With MEC,[1] many applications and services will benefit from localized communication, storage, processing, and management, thus dramatically decreasing service response latency, reducing the traffic load on the core network, and improving context awareness [5]. The MEC concept is not in collision with MCC; they complement each other in building flexible and reconfigurable 5G networks using a "network slicing" approach, where different services may easily be instantiated using different virtualized architectures on top of the high-performance MEC host nodes. Instrumental to the development of flexible packet core networks are novel 5G reconfigurable architectures based on software defined networking (SDN) and network functions virtualization (NFV) [6]. Thus, enhanced with support for MEC, 5G mobile core networks will provide a distributed information processing and storage architecture that is ideally suited to services requiring low latency and localized decision making.

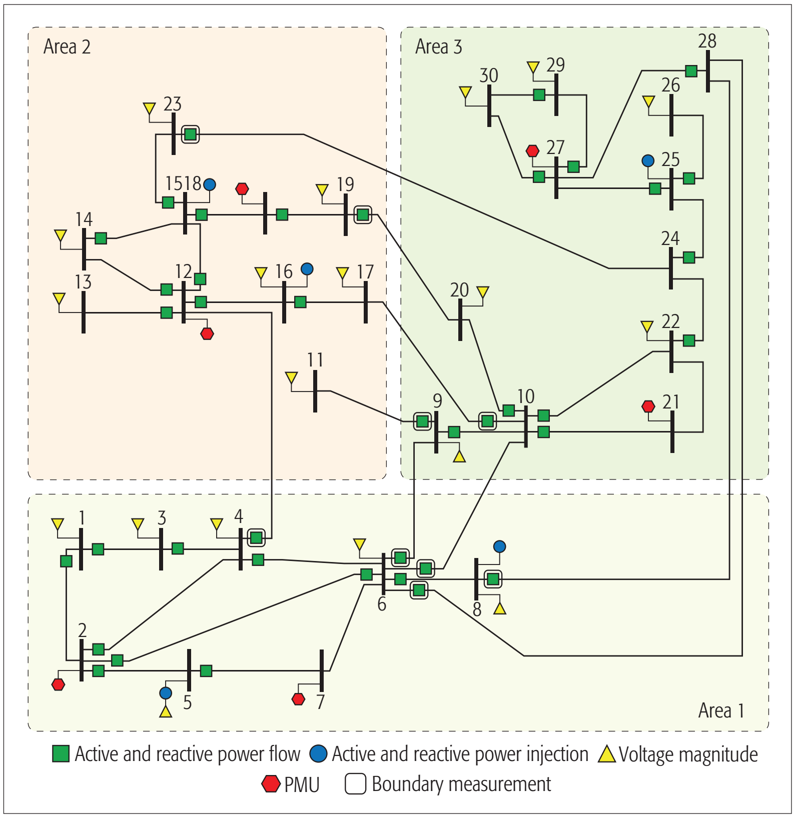


Figure 1. The IEEE 30 bus test case segmented into three areas with a given collection of measurements.

Distributed State Estimation Methods

The deregulation of energy markets, necessary for real-time monitoring and control, along with utilities' data security and privacy concerns in multi-area settings substantiate the need for developing *hierarchical* and *distributed* SE methods as an alternative to classical *centralized* schemes.

[1] With a slight abuse of notation, hereinafter we use the term "MEC." In some passages, however, "fog computing" could be deemed more appropriate.

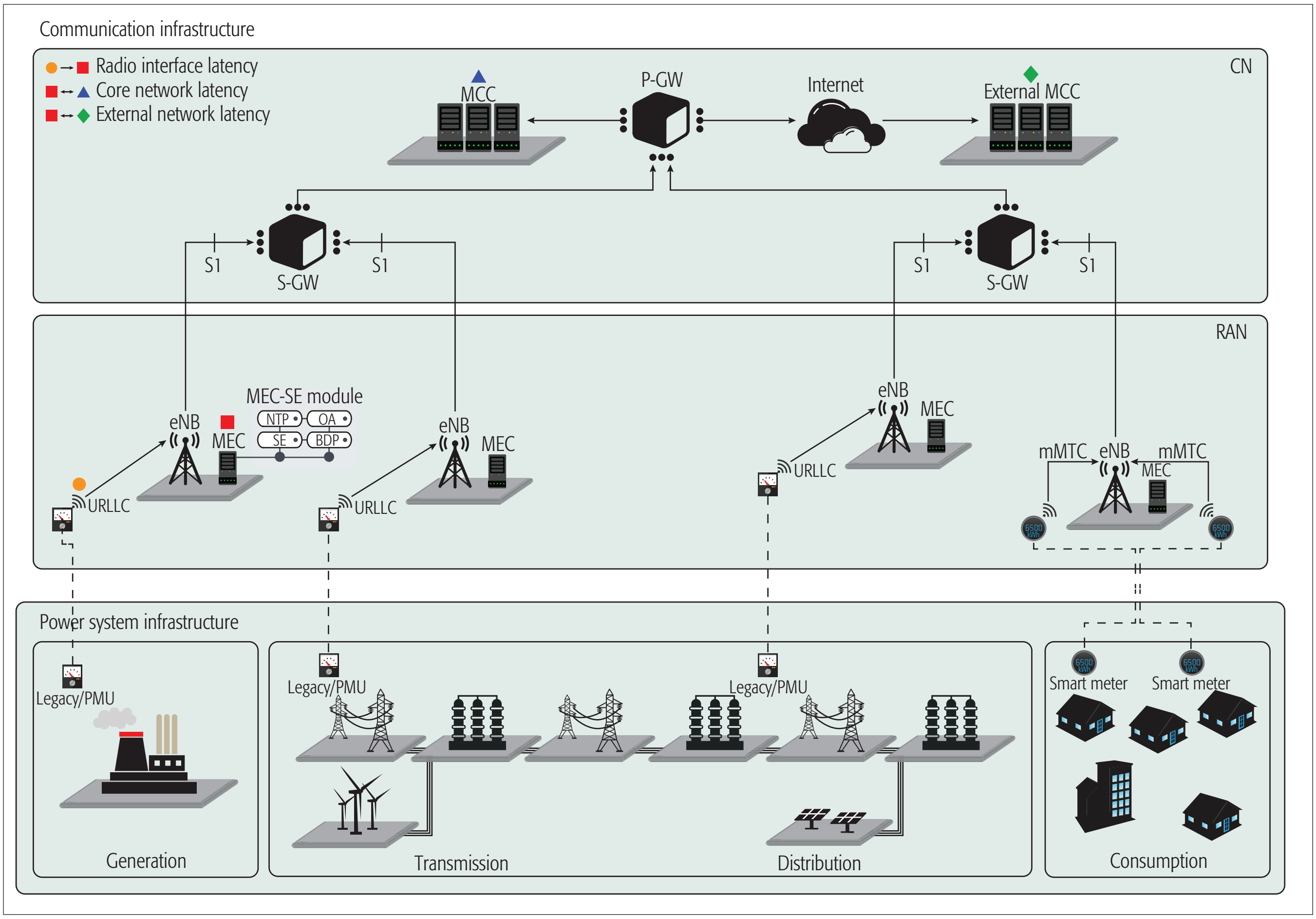


Figure 2. The architecture with two layers: i) power system infrastructure and ii) communication infrastructure that combines novel RAN interfaces supporting mMTC and URLLC, and new virtualized core network (CN) MEC/MCC-based architecture with network topology processor (NTP), observability analysis (OA), state estimation algorithm (SE), and bad data processing (BDP) routines to support future smart grid services such as distributed SE.

Hierarchical and Distributed SE Methods

In *hierarchical SE* [1, 7], a central authority controls the local processor in each area or level. Gomez-Exposito *et al.* [1] propose a hierarchical multi-level SE scheme where local estimates are computed at lower voltage levels and transferred to higher voltage areas, up to the system operator level, in order to estimate the system-wide state. In each stage, the SE problem is solved via the Gauss-Newton method. Still in a hierarchical context, Korres in [7] proposes to decompose, on a geographical basis, the overall system into a number of subsystems in non-overlapping areas. Each area independently runs its own gradient-based SE scheme on the basis of local measurements. Such estimates are then communicated to the central coordinator, which computes the system-wide solution.

In *distributed* approaches [8–10], on the contrary, each local processor communicates only with its neighbors, since no central authority exists. The authors in [8] propose a distributed SE scheme based on primal-dual decomposition. This method requires the exchange of information only between *neighboring* areas, namely, border state variables and the dual variables. For each area, the problem is solved through classical Gauss-Newton techniques. Differently, Kekatos and Giannakis [9] resort to the ADMM to solve the SE problem in a distributed fashion. In contrast to [8], the authors develop a robust version leveraging on the sparsity of bad data measurements. Going one step beyond, [11] proposes a *hybrid* scheme including *both* PMUs and legacy measurements. Here, the SE problem is cast into a semidefinite programming framework and solved via convex semidefinite relaxation techniques, in both centralized and decentralized settings. In [10], the authors propose a hybrid multi-area state estimator based on successive convex approximation (SCA) and ADMM. The proposed distributed approach is equivalent to the centralized case in terms of estimation accuracy and is able to operate in broader scenarios where the semidefinite relaxation approach fails.

Going one step beyond, other authors [12, 13], have considered *fully distributed* SE approaches where interaction takes place at the bus level rather than the area level. Hu *et al.* [12] pioneered in the application of a message-passing BP algorithm to the SE problem, where the system state is modeled as a set of stochastic variables. This provides a flexible solution for the inclusion in the model of distributed power sources, environmental correlation via historical data and time-varying loads, and so on. In a recent work, a distributed Gauss-Newton algorithm based on factor graphs and a BP algorithm is proposed and shown to provide the same accuracy as the centralized Gauss-Newton algorithm [13], while being flexible enough to accommodate both fully distributed and multi-area SE scenarios.

As representative methods, ADMM and BP are particularly promising, and thus are described with further detail later in this section. Prior to that, we describe a system model suitable for both approaches.

System Model

The SE aims to determine the values of the state variables based on knowledge of the network topology and measurements obtained from devices spread across the power system. Thus, the SE problem reduces to solving the system of equations: $\mathbf{z} = \mathbf{h}(\mathbf{x}) + \mathbf{u}$, where $\mathbf{h}(\mathbf{x}) = (h_1(\mathbf{x}), ..., h_k(\mathbf{x}))$ may include both nonlinear (from legacy metering devices) and linear measurement functions (from PMUs); $\mathbf{x} = (x_1, ..., x_n)$ is the vector of the state variables; $\mathbf{z} = (z_1, ..., z_k)$ is the vector of independent measurements (where $n < k$); and $\mathbf{u} = (u_1, ..., u_k)$ is the vector of measurement errors. The state variables are bus voltage magnitudes and bus voltage angles, along with transformer magnitudes of turns ratio and transformer angles of turns ratio. Figure 1 below illustrates a possible scenario for the collection of measurements in the IEEE 30 bus test case.

Optimization-Based Distributed SE Methods

ADMM is experiencing renewed popularity after its discovery in the mid-20th century. ADMM was conceived to overcome the weaknesses of its predecessors: the primal-dual decomposition method and the method of multipliers. The former is suitable for distributed optimization but presents convergence issues for non-differentiable objectives. Conversely, the method of multipliers can deal with non-smooth functions but couples the objective function, which makes it barely suitable for distributed optimization. ADMM brings the two features together: it is suitable for distributed implementation and can efficiently deal with non-differentiable objective functions.

The canonical optimization problem solved by ADMM is the minimization of a composite objective function (i.e., $f(\mathbf{x})+ g(\mathbf{z})$) subject to a linear equality constraint of the form $\mathbf{Ax} + \mathbf{Bz} = \mathbf{c}$, with $\mathbf{x}$ and $\mathbf{z}$ being the optimization variables. To deal with non-differentiable functions, ADMM augments the cost function by a quadratic penalty term that transforms the optimization problem into a strongly convex problem but with the same stationary solution. This transformation has major implications in the dual domain, as the dual function becomes differentiable. Then ADMM iterates sequentially on the primal optimization variables, $\mathbf{x}$ and $\mathbf{z}$, and the dual variables until convergence.

To decentralize an optimization problem, ADMM decouples the objective function with consensus variables. Consensus variables introduce equality constraints into the optimization problem, separating it into a number of subproblems [9, 10]. The resulting ADMM algorithm can be interpreted as an iterative message passing procedure, in which the agents solving the subproblems (e.g., utilities in the multi-area SE problem) exchange consensus and dual variables until convergence. Besides, ADMM can be used in conjunction with SCA approaches that efficiently deal with non-convex problems.

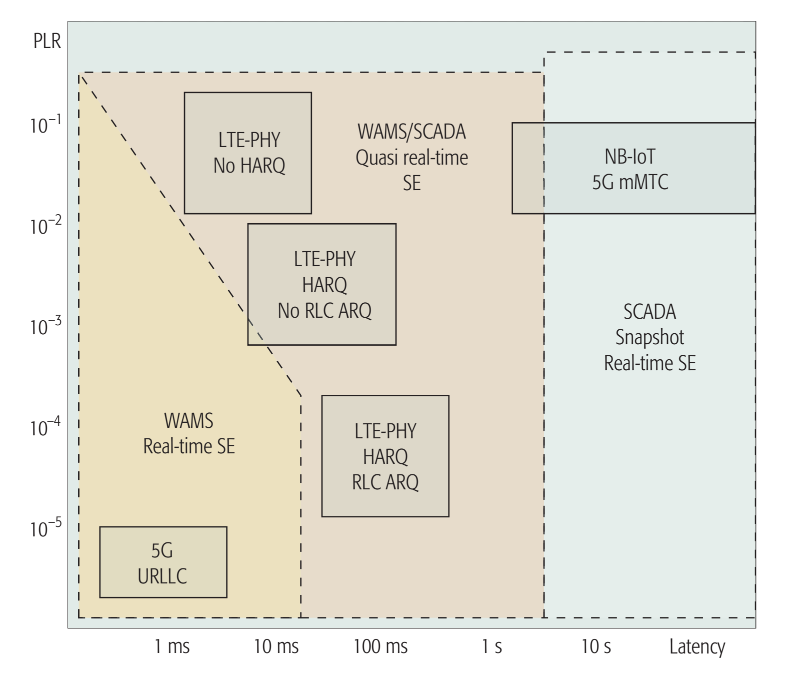


Figure 3. Reliability and latency performance of 4G/5G radio interface solutions relative to different WAMS/SCADA SE services.

Probabilistic Inference-Based Distributed SE Methods

Probabilistic graphical models, such as factor graphs, provide a convenient framework to represent dependencies among the system of random variables, such as the state variables $\mathbf{x}$ of the power system. Specifically, the bus/branch model with a given measurement configuration is mapped onto an equivalent factor graph containing the set of *factor* and *variable* nodes. *Factor nodes* are defined by the set of measurements: arbitrary factor node *f* is associated with measured value *z*, measurement error *U*, and measurement function $h(\mathbf{x})$. *Variable nodes* are determined by the set of state variables $\mathbf{x}$. A factor node is connected by an edge to a variable node if and only if the state variable is an argument of the corresponding measurement function $h(\mathbf{x})$.

When applied on factor graphs, BP algorithms allow marginal distributions of the system of random variables to be calculated efficiently. BP is a distributed message-passing algorithm in which two types of messages are exchanged along the edges of the factor graph: messages from a variable node to a factor node and vice versa. In general, for SE scenarios a *loopy* BP algorithm must be used since the corresponding factor graph contains cycles. Loopy BP is an iterative algorithm in which, for the standard scheduling, messages are updated in parallel in respective half-iterations. Within half-iterations, factor nodes calculate and send messages to incident variable nodes, while subsequently, variable nodes calculate and send messages back toward factor nodes. As a general rule, an output message on any edge exclusively depends on incoming messages from all other edges. BP messages represent beliefs about variable nodes; thus, a message that arrives or departs

Focusing only on data-plane elements, the packet core network consists of packet gateways (S-GW/P-GW) that interconnect base stations and provide access to external networks (e.g., the Internet). Base stations connect to core network gateways via S1 interfaces, and may also be directly interconnected via X2 interfaces.

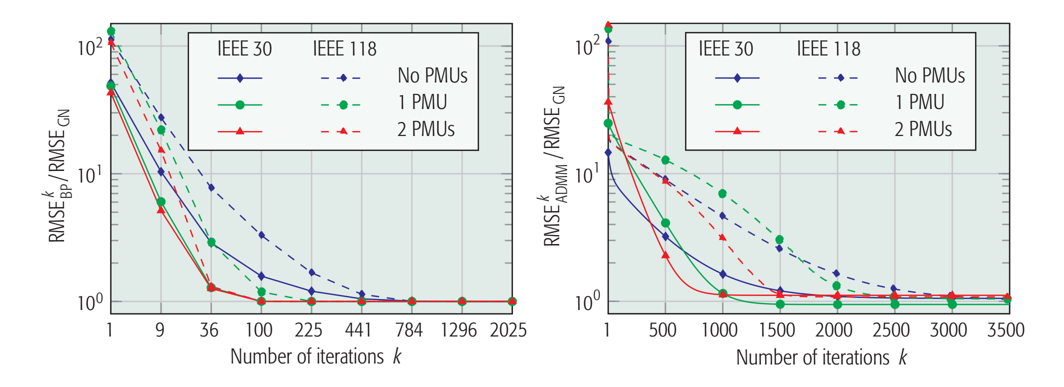


Figure 4. Normalized RMSE for the BP algorithm (subfigure a) and ADMM (subfigure b) in three scenarios: without PMUs, one PMU and two PMUs per area for the IEEE 30 and IEEE 118 bus test case.

from a certain variable node is a function (distribution) of the random variable corresponding to the variable node. Finally, the marginal inference provides an estimate of the state variables (voltages in the power system).

Proposed 5G System Architecture for Distributed Smart Grid Services

In this section, we propose to leverage an emerging 5G network architecture, in particular, its features outlined earlier, to enable deployment of advanced smart grid services such as distributed SE. We also discuss the latency and reliability requirements that distributed SE imposes on 5G communication networks.

System Architecture

In Fig. 2, we present the proposed system architecture for distributed smart grid services. The lowermost layer represents the electrical grid broken down into the generation, transmission, distribution, and consumption network segments. The grid is equipped with a large number of measurement devices ranging from legacy RTUs to PMUs and massive-scale smart meter infrastructure. We assume the grid is organized into a number of non-overlapping areas. Such a multi-area SE problem represents an input to the distributed SE algorithms discussed earlier.

As far as the communication technology is concerned, hereinafter we focus on 5G networks. Still, LTE is used as a reference where needed. The electrical grid is covered by the RAN comprising a large number of base stations (eNBs). Focusing only on data plane elements, the packet core network consists of service and packet gateways (S-GWs/P-GWs) that interconnect eNBs and provide access to external networks (e.g., the Internet). The eNBs connect to core network gateways via S1 interfaces, and may also be directly interconnected via X2 interfaces.

The support for MCC/MEC within the packet core network is provided in the form of a data center for MCC, and in the form of a large-scale deployment of smaller data and computing centers in the vicinity of eNBs at the network edge (for MEC). MEC nodes host the distributed smart grid applications (SE, topology processor, etc.; Fig. 2). In the sequel, we focus on distributed SE modules denoted as MEC-SE modules. Using NFV concepts, MEC-SE modules may run within the virtualization environment of MEC nodes, that is, they can be remotely instantiated, removed, and orchestrated using the centralized NFV orchestrator.

Connectivity between remote measurement devices, MEC-SE modules, and the MCC-SE module is provided by a 5G network. For connections between measurement devices and local MEC-SE modules, 5G will offer flexible wireless interfaces for different measurement devices. Massive-scale smart meters will upload their data via mMTC service, while more stringent reliability and latency can be offered to RTUs and PMUs via URLLC service. We assume the smart metering data will be delivered as aggregated measurements using data aggregation units. Data flows between MEC-SE modules can be flexibly established via S1 or X2 interfaces. SDN-based concepts could be applied to, say, connect the distributed MEC-SE modules to the central MCC-SE module. This module may or may not participate in the distributed SE process and also serve as a central function and data repository interfacing other energy management functions.

Latency and Reliability Requirements for Distributed SE

Utilizing PMU inputs, WAMS will enable power system operators to monitor their power networks in real time. PMUs track system state variables (phasors) with sampling rates of 10–20 ms. Targeting ever faster reaction to system disturbance, decentralized WAMS architecture employing distributed SE with localized decision making promises minimal system response latency. In the following, we discuss how well 5G reliability and latency targets match the vision of future real-time WAMS.

Consider MEC-SE modules, each running an entity of the distributed SE algorithm whose scope is the surrounding geographic area. Every MEC-SE module continuously updates its local state estimates based on high-rate PMU inputs additionally supported by legacy RTU measurements. Furthermore, neighboring MEC-SE modules exchange messages to execute the distributed SE algorithm, thus further refining their state estimates. Clearly, the distributed SE performance critically depends on the latency and reliability of the communication links:

- From a PMU to MEC-SE module
- Between two neighboring MEC-SE modules

Note that the latency and reliability of MEC-SE to MCC-SE module communication does not affect distributed SE, but it would affect both the hierarchical and centralized SE.

Figure 3 provides an overview of different 4G/5G radio interface solutions that affect the reliability/latency trade-off of the link between a measurement device and an MEC-SE module. From the figure, it is evident that the current 4G LTE radio interface imposes reliability/latency trade-off limits that prevent the real-time WAMS goals of tracking the system state with latencies as low as PMUs sampling rates. As discussed in [14] and empirically investigated in [15], the LTE PHY interface latency may be decreased to 15–20 ms in the uplink (due to uplink scheduling requests/grants), and down to 4 ms in the downlink if both medium access control (MAC) layer hybrid automatic repeat request (HARQ) mechanism are avoided, but with modest PLR ~ 10^{-1}. The PLR may be gradually decreased down ~ 10^{-5} by including up to three HARQ retransmissions, and by using RLC ARQ, at the price of increased latency to ~ 40–60 ms, thus allowing only quasi real-time SE. In contrast, 5G URLLC fits the future real-time WAMS targets with radio interface latency of ~ 1 ms and PLRs 10^{-5}. We note that low-cost interfaces such as NB-IoT/5G mMTC could serve massive AMI connections, as well as the needs of legacy SCADA-based snapshot SE.

The latency within packet core networks is very low, typically in the range of 3–10 ms between MEC-SE modules, and between MEC-SE and MCC-SE modules (for moderately large networks). Note that in the case of two MEC-SE modules residing at two eNBs connected via X2 interface, this latency can be reduced to ~ 1 ms. In contrast, for scenarios where the central MCC-SE module is hosted in an external data center outside the core network, the associated latency can be as high as 10–100 ms. In the following section, we place ADMM and BP-based distributed SE in the context of 5G system architecture, investigating the impact of latency and reliability on the SE performance.

Performance of Distributed State Estimation Methods

Both BP and ADMM-based distributed SE solutions can be integrated as part of the 5G smart grid services described in the previous section. For the case of BP, factor graphs of the power system can be flexibly segmented into areas, and BP can easily accommodate both intra- and inter-area message exchange, not necessarily with the same periodicity, allowing for asynchronous message scheduling. Thus, a factor graph of each area can be maintained within the corresponding MEC-SE module, with local measurements arriving from mMTC and URLLC connections. Inter-area BP messages can be exchanged with a controlled periodicity between the neighboring MEC-SE modules. The exchange of inter-area BP messages establishes a global factor graph and provides the MEC-SE modules with the ability to converge to the global solution. Similarly, for the case of ADMM, the local ADMM-based MEC-SE modules may run single-area optimization based on local topology and measurements. By exchanging messages among neighboring areas, MEC-SE modules iterate through the ADMM optimization process converging toward the global solution.

Performance of BP and ADMM-Based SE

In the following, we demonstrate that the state estimate of the distributed BP and ADMM-based algorithms converges to the solution provided by the centralized Gauss-Newton method. We consider both the IEEE 30 and IEEE 118 bus test cases divided into three and nine areas, respectively (Fig.1 for IEEE 30 bus test case). The algorithms are tested in three scenarios:

- A measurement configuration with no PMUs
- One PMU per area
- Two PMUs per area

For a predefined value of the noise variance and using a Monte Carlo approach, we generate 500 random sets of measurement values, feed them to the BP, ADMM, and centralized SE algorithms, and then compute the average performance. The BP algorithm is implemented as a BP-based distributed Gauss-Newton method described in [13], which can be interpreted as a fully distributed Gauss-Newton method. The ADMM-based algorithm is based on [10], where the SCA scheme in the outer loop is combined in a distributed fashion within the iterative framework of ADMM, which constitutes the inner loop. To evaluate both algorithms, we use the root mean square error (RMSE) after each iteration k (RMSE^k), normalized by the RMSE of the centralized SE algorithm using the Gauss-Newton method after 12 iterations ($\mathrm{RMSE}_{\mathrm{GN}}$).

Figure 4a shows that the BP algorithm converges to the solution of the centralized SE for each scenario. As expected, the BP algorithm converges faster for measurement configurations with PMUs. In general, configurations with PMUs can dramatically improve numerical stability of the BP algorithm and prevent oscillatory behavior of messages. Figure 4b illustrates the performance of the ADMM-based algorithm. The scheme attains the same performance as the centralized SE. We observe how an increased number of PMUs leads to significant improvement in convergence behavior for both the IEEE 30 and IEEE 118 bus test cases. For the latter, the graph reveals that the algorithm needs a larger number of iterations to converge. Note that in both cases, nonlinear measurement functions (SCADA) are used, and the algorithm is initialized in a (flat-start) state distant from the solution, allowing only snapshot SE with order of seconds latency.

Distributed SE Methods: Performance vs. Reliability

In Fig. 5, we illustrate the RMSE performance of the ADMM-based scheme as a function of PLR. We consider the IEEE 30 bus test case scenario with PMU measurements only, with guaranteed observability, and corrupted with additive white Gaussian noise of standard deviation $\sigma = 10^{-4}$. Whenever a measurement (packet) is dropped, it is replaced by a pseudo-measurement with higher standard deviation σ_{pm} (Fig. 5). The PLR of interest is in the range of 10^{-5} (for 5G URLLC service) and 10^{-1} (for LTE without HARQ or RLC mechanisms in order to meet latency requirements). The performance of a centralized SE scheme based on the Gauss-Newton approach is also presented [1].

The exchange of inter-area BP messages establishes a global factor graph and provides the MEC-SE modules with the ability to converge to the global solution. Similarly, for the case of ADMM, the local ADMM-based MEC-SE modules may run single-area optimization based on local topology and measurements.

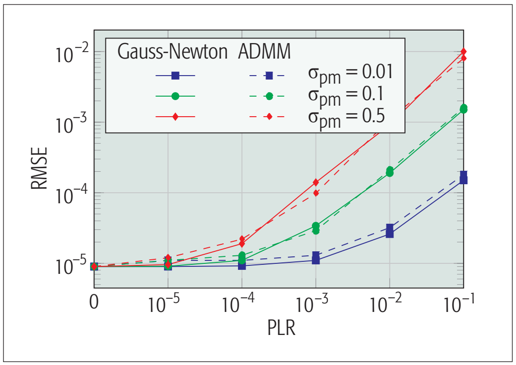


Figure 5. RMSE vs. packet loss rate for the IEEE 30 bus test case with full PMU observability.

The performance attained by the distributed ADMM scheme for PLR of 10^{-5} (URLLC) is very close to that in the absence of packet losses. As expected, performance severely degrades by up to three orders of magnitude when PLR increases from 10^{-5} to 10^{-1} (i.e., in LTE range). The degradation is comparable for both the ADMM-based and Gauss-Newton approaches; however, ADMM operates in a distributed manner, providing better scalability while preserving privacy.

Distributed SE Methods: Performance vs Latency

In Fig. 6, we analyze the latency of the BP-based scheme in three different scenarios for the IEEE 30 bus test case. We analyze the scenario where the system changes both generation and load values at the time instant t = 10 ms. As an example, we focus on the sudden bus voltage magnitude drop V_3. We assume measurements are obtained synchronously and immediately after the system change at t = 10 ms using PMUs only, thus resulting in the linear SE model. The BP-based SE model runs the iterative message-passing algorithm continuously over time, with new measurements being integrated as they arrive.

Figure 6a illustrates the BP-based SE scenario, which ignores communication latencies, while focusing only on computational latency measured from the time instant when the system acquired a new set of PMU measurements. The BP-based algorithm is able to provide fast response on the new state of the power system, and steady state occurs after several BP iterations (note that figure markers denote BP outputs after each iteration). The computational latency can be additionally reduced if BP computations across variable/factor nodes in every iteration are parallelized.

Next, we assume PMUs deliver their measurements to MEC nodes through URLLC connection, which introduces latency of 2 ms. In addition, we consider the behavior of the distributed and asynchronous BP-based SE. More precisely, local MEC-SE modules run BP in a distributed fashion, where neighboring areas (MEC-SE modules) asynchronously exchange messages via X2 interfaces that introduce latency of 1 ms. Figure 6b shows that the BP algorithm requires more iterations to reach the steady state due to delay of updates of inter-area BP messages, which results in computational latency increase. However, the BP algorithm is still able to track the system changes at the level of ~ 10 ms.

Finally, we consider the scenario where PMU measurements are forwarded to the MCC node, where we observe two cases:

- The MCC node resides within the mobile core network.
- The MCC node is part of a data center in some external network.

For the former case, typical core network latency of 10 ms is considered, while for the latter, additional latency of 20 ms toward the external network is assumed. In both cases, the BP algorithm is implemented in the MCC node. In this framework, the SE model provides quasi real-time performance (Fig. 3). We also note that if the 4G LTE interface is used instead of 5G URLLC, the additional latency of ~ 20–40 ms needs to be included.

Conclusions

In this article, we present convincing evidence that in forthcoming years, 5G technology will provide an ideal arena for the development of future distributed smart grid services. These services will rely on massive and reliable acquisition of timely information from the system, in combination with large-scale computing and storage capabilities, providing a highly responsive, robust, and scalable monitoring and control solution for future smart grids.

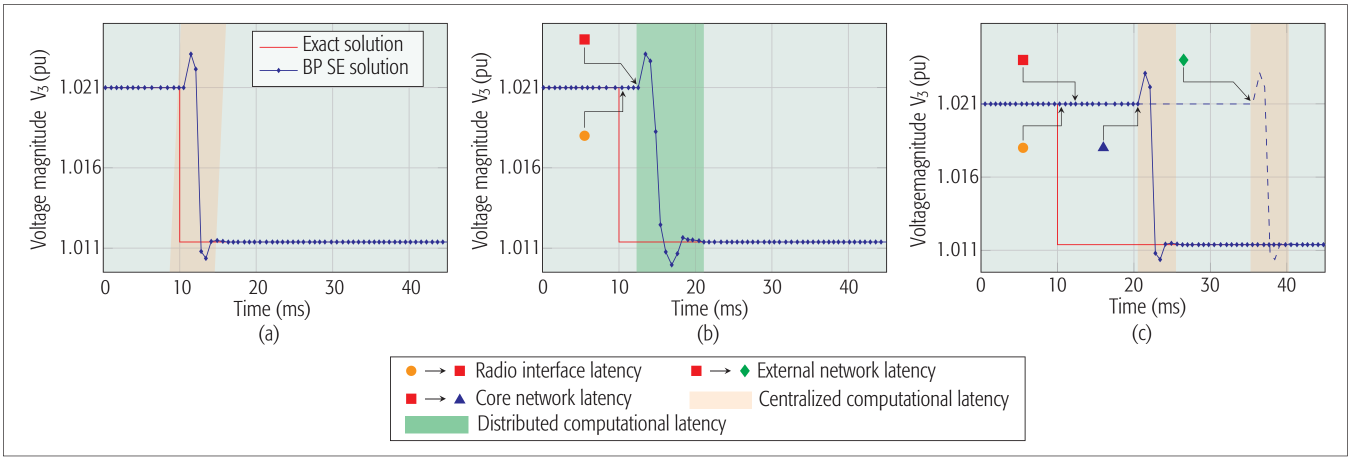


Figure 6. Latency performance of the BP-based SE for the IEEE 30 bus test case where the BP-based SE: a) runs in a mode where communication latencies are neglected; b) runs in a distributed mode accounting for radio interface and MEC-to-MEC node latencies; c) runs in the MCC node with radio interface and core network latency (dash) or external network latency (solid).

Acknowledgment

This article has received funding from the EU 7th Framework Programme for research, technological development, and demonstration under grant agreement no. 607774.

References

[1] A. Gomez-Exposito *et al.*, "A Multilevel State Estimation Paradigm for Smart Grids," *Proc. IEEE*, vol. 99, no. 6, June 2011, pp. 952–76.

[2] H. Shariatmadari *et al.*, "Machine-Type Communications: Current Status and Future Perspectives Toward 5G Systems," *IEEE Commun. Mag.*, vol. 53, no. 9, Sept. 2015, pp. 10–17.

[3] Y. C. Hu *et al.*, "Mobile Edge Computing — A Key Technology Towards 5G," ETSI white paper, vol. 11, 2015.

[4] A. Rico-Alvarino *et al.*, "An Overview of 3GPP Enhancements on Machine to Machine Communications," *IEEE Commun. Mag.*, vol. 54, no. 6, June 2016, pp. 14–21.

[5] F. Bonomi *et al.*, "Fog Computing and Its Role in the Internet of Things," *Proc. 1st MCC Wksp. Mobile Cloud Computing*, 2012, pp. 13–16.

[6] A. Maeder *et al.*, "A Scalable and Flexible Radio Access Network Architecture for Fifth Generation Mobile Networks," *IEEE Commun. Mag.*, vol. 54, no. 11, Nov. 2016, pp. 16–23.

[7] G. N. Korres, "A Distributed Multiarea State Estimation," *IEEE Trans. Power Systems*, vol. 26, no. 1, Feb. 2011, pp. 73–84.

[8] E. Caro, A. J. Conejo, and R. Minguez, "Decentralized State Estimation and Bad Measurement Identification: An Efficient Lagrangian Relaxation Approach," *IEEE Trans. Power Systems*, vol. 26, no. 4, Nov. 2011, pp. 2500–08.

[9] V. Kekatos and G. B. Giannakis, "Distributed Robust Power System State Estimation," *IEEE Trans. Power Systems*, vol. 28, no. 2, May 2013, pp. 1617–26.

[10] J. Matamoros *et al.*, "Multiarea State Estimation with Legacy and Synchronized Measurements," *IEEE ICC 2016*, May 2016, pp. 1–6.

[11] H. Zhu and G. B. Giannakis, "Power System Nonlinear State Estimation Using Distributed Semidefinite Programming," *IEEE J. Selected Topics in Signal Processing*, vol. 8, no. 6, Dec. 2014, pp. 1039–50.

[12] Y. Hu *et al.*, "A Belief Propagation Based Power Distribution System State Estimator," *IEEE Computational Intelligence Mag.*, vol. 6, no. 3, Aug. 2011, pp. 36–46.

[13] M. Cosovic and D. Vukobratovic, "Distributed Gauss-Newton Method for AC State Estimation: A Belief Propagation Approach," *2016 IEEE Int'l. Conf. Smart Grid Commun.*, Nov. 2016, extended version: https://arxiv.org/abs/1702.05781, pp. 643–49.

[14] A. Larmo *et al.*, "The LTE Link-Layer Design," *IEEE Commun. Mag.*, vol. 47, no. 4, Apr. 2009, pp. 52–59.

[15] M. Laner *et al.*, "A Comparison Between One-Way Delays in Operating HSPA and LTE Networks," *2012 10th Int'l. Symp. Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, May 2012, pp. 286–92.

Biographies

Mirsad Cosovic (mirsad.cosovic@schneider-electric-dms.com) received his Dipl.-Ing, and Mr.-Ing. degrees in power electrical engineering from the University of Sarajevo, Faculty of Electrical Engineering, Bosnia and Herzegovina, in 2009 and 2013, respectively. Since December 2009, he has been a teaching assistant in the Faculty of Electrical Engineering of Sarajevo, and since October 2014, he has been a Ph.D. candidate as a Marie Curie Early Stage Researcher at the University of Novi Sad, Faculty of Technical Sciences, Serbia.

Achilleas Tsitsimelis (achilleas.tsitsimelis) received his Dipl.-Ing in electrical and computer engineering in 2013 from the National Technical University of Athens (NTUA). During 2013 and 2014, he worked as a researcher at the Electrical and Computer Engineering School, NTUA-ICCS. Since 2014 he has been a Marie Curie Early Stage Researcher at CTTC, and he is pursuing his Ph.D in signal theory and communications at the Universidad Politecnica de Catalunya (UPC). His research interests include state estimation for the smart grid.

Dejan Vukobratovic (dejanv@uns.ac.rs) received his Ph.D. degree in electrical engineering from the University of Novi Sad in 2008. In 2009 he became an assistant professor and in 2014 an associate professor at the Department of Power, Electronics and Communication Engineering, University of Novi Sad. During 2009 and 2010, he was a Marie Curie Intra-European Fellow at the University of Strathclyde, United Kingdom. His research interests include information and coding theory, wireless communications, and signal processing.

Javier Matamoros (javier.matamoros) holds a researcher position at CTTC. He received his M.Sc. degree in telecommunications and Ph.D. degree in signal theory and communications from UPC in 2005 and 2010, respectively. He has participated in several national and EC-funded projects (JUNTOS, NEWCOM#, E2SG, EXALTED, ADVANTAGE). His primary research interests are distributed optimization and machine learning applied to communications and smart grids.

Carles Antón-Haro [M'99, SM'03] (carles.anton@cttc.es) holds a Ph.D. degree in telecommunications from UPC (cum laude). In 1999, he joined Ericsson, where he participated in rollout projects of 2G/3G networks. Currently, he is with the CTTC as the director of R&D Programs and a senior researcher. His research interests include signal processing for communications (MIMO, WSN, 5G) and smart grids, optimization, estimation, and control. He has published +30 articles in IEEE journals and +100 conference papers

5G technology will provide ideal arena for the development of future distributed Smart Grid services. These services will rely on massive and reliable acquisition of timely information from the system, in combination with large-scale computing and storage capabilities, providing highly responsive, robust and scalable monitoring and control solution for future Smart Grids.

Energy Big Data Security Threats in IoT-Based Smart Grid Communications

Wen-Long Chin, Wan Li, and Hsiao-Hwa Chen

To deal with security threats, energy big data should be thoughtfully stored and processed to extract critical information, and security and blackout warnings should be given in an early stage. The authors present a comprehensive tutorial and survey to highlight research challenges related to these issues in the IoT-based smart grid.

Abstract

Increased intelligence and automation in smart grid results in many heterogeneous applications benefiting from the Internet of Things, such as demand response, energy delivery efficiency/reliability, and fault recovery. However, vulnerabilities in smart grid arise due to public communication infrastructure and Internet-based protocols. To deal with security threats, energy big data should be thoughtfully stored and processed to extract critical information, and security and blackout warnings should be given in an early stage. This work gives a comprehensive tutorial and survey to highlight research challenges on the aforementioned issues in the Internet-of-Things-based smart grid. We demonstrate that a stealthy and blind energy big data attack can be launched using a replay scheme. Also, we elucidate an intuitive geometric viewpoint for this type of attack. The proposed attack can bypass bad data detection successfully using either DC or AC state estimation.

Introduction

The Internet of Things (IoT) is an emerging technology expected to change our daily life rapidly. Via interworking of different devices, any physical object/thing can be integrated seamlessly for exchanging and collecting data. Objects in the physical world, including fridges, heaters, televisions, and so on, could be easily accessible and manageable. The IoT allows devices to be sensed and controlled remotely across existing networks, resulting in improved efficiency and economic benefits. With the IoT technology, smart grid (SG) becomes an instance of cyber-physical systems [1, 2] The development of most parts of SG can be enhanced by applying IoT. Through IoT, the whole power grid chain, from electricity generation to consumption, will enable intelligence and two-way communication capabilities to monitor and control the power grid anywhere and anytime.

While the IoT technology is very important in the context of SG, it could also lead to disasters since the operations of SG are based on Internet-based protocols [3, 4]. Therefore, the utility is exposed to general information and communication technology (ICT) threats, such as denial of service (DoS) attacks and domain-specific attacks (e.g., targeted malware such as Stuxnet). As a consequence, an attacker could create huge financial losses and damages to the utility by inducing real-time imbalance between energy consumption and generation through data manipulation. If the operators cannot locate the vulnerabilities of the SG rapidly and accurately, it is easy to trigger serious events leading to a breakdown of a power grid. Therefore, secure, reliable, and real-time situational awareness is critical for future power grids.

The pervasive deployment of smart metering in IoT-based SG will generate energy big data in terms of its huge volume, large scale, and structural variety. Three categories of grid business data are listed as follows [5]:

1. Grid operation and equipment testing or monitoring data, such as supervisory control and data acquisition (SCADA) data and sampling data of smart meters
2. Electric power marketing data, such as transaction price and electricity sales data
3. Electric power management data, such as internal grid data

These data must be processed in a parallel and distributed fashion to extract critical information for decision making processes within a limited time. According to inherent data structure, energy data are divided into structured and unstructured data. Structured data includes data mainly stored in relational databases.

The growth rate of structured data is extremely high. Big data will lead to the challenges in distributed storage of power systems, and a distributed storage system for managing structured data that is designed to scale to a very large size is desirable. There are many potential advantages to be derived from energy data for the goal of optimal operation, including real-time monitoring of energy consumption data generated by advanced metering infrastructure (AMI) and smart meters, detection of energy losses by fault or fraud, early blackout warning, fast detection of disturbances in energy supply, and intelligent energy generation, planning, and pricing. Huge data generated at the second level and concurrent peak demands from different homes may cause blackouts at some substations due to power imbalance introduced by inaccurate energy forecast. Energy big data are also very useful for realizing situational awareness. Based on long-term monitoring, security-related information can also be characterized.

In this work, we demonstrate a stealthy and blind energy big data attack using a replay mechanism without requiring the information of power grid topology and transmission-line admittances. In contrast to conventional data falsification attacks using cumbersome mathematical approaches, we elucidate an intuitive geometric approach for this

This work was supported in part by the Ministry of Science and Technology of Taiwan (105-2221-E-006-019-MY2 and 104-2221-E-006-081-MY2).

Digital Object Identifier:
10.1109/MCOM.2017.1700154

The authors with National Cheng Kung University.

type of attack. The proposed attack can bypass bad data detection (BDD) successfully via either direct current (DC) or alternating current (AC) state estimation. Interesting readers can refer to [6] for more information about the state estimation.

The major contributions of this work are outlined as follows.

- We survey the security and energy big data analytics issues of IoT-based SGs. The potential applications of energy big data analytics are also introduced. Future research challenges are outlined for the IoT-based SG applications.
- We demonstrate a new energy big data attack employing a replay approach with both DC and AC state estimations. To the best of our knowledge, no works have been done to study data falsification attacks using an AC power flow model. No reports have appeared on launching blind AC attacks without grid parameters, such as transmission-line admittances.
- The effectiveness of the proposed big data attack is verified by simulations.

The rest of this article can be outlined as follows. Security and energy big data analytics issues are discussed. We illustrate the proposed big data attack, and we evaluate the performance and vulnerability of IoT-based SG. We highlight future research challenges. Finally, we draw the conclusions of this article.

Security and Energy Big Data Analytics Issues

IoT-Based Smart Grid Security Issues

Security in critical utility infrastructure is a very serious concern and involves many factors, including physical security of plants and facilities, SCADA, intelligent electronic devices (IEDs) and meters, cyber security for networking and computing, and security management for the utility. The SG will encompass billions of smart objects via IoT networks, including smart meters, smart appliances, sensors, actuators, and so on. However, there have been a lot of concerns regarding vulnerabilities of the SG. The security threats outlined below are the major factors impeding rapid and wide deployment of the IoT-based SG [7–9].

Impersonation: The attacker acts on behalf of a legitimate user in an unauthorized way. To solve this problem, a framework of machine-to-machine authentication in SG via a two-layer approach was proposed in [10].

Eavesdropping: Since the IoT uses public communication networks, an attacker can easily intercept the energy consumption information of households.

Data manipulation: Modifying exchanged data may cause service impairment threats, such as DoS, compromise of service, and corruption of energy data. Recently, a DC blind false data attack [11] was reported without knowing power grid topology and transmission-line admittances.

Access and authorization: Distributed devices can be accessed and controlled remotely. Meters and other devices can be compromised by malicious software codes. The infiltration threat relates to the penetration of a secure perimeter by an unauthorized access, and can allow other threats to be exercised.

Availability: Large-scale IoT-based SGs are vulnerable to IP-based attackers, making them partially or totally unavailable as a result of DoS attacks [12].

Energy Big Data Analytics Issues in IoT-Based Smart Grid

Upgrading utility networks will force electricity providers to process far more information than ever before [13]. To make full use of the new data, the utility companies will need complex event-processing capabilities as listed below.

Scalable, interoperable, and distributed computing infrastructure: As SG is a highly distributed system, a huge amount of data is collected from every section, including energy generation, transmission, distribution, and renewal energy powered vehicles and smart meters. It is very challenging to store, share, and process such volume, velocity, and variety (3V) big data.

Real-time big data intelligence: Real-time decision is essential for both system operation and real-time pricing. Intelligent decision making needs to process current and past data. With the real-time constraints, it will be extremely challenging to design new algorithms that can provide intelligence for processing such big data.

Big data knowledge representation and processing: Big data analytics requires new machine learning and artificial intelligence theories. However, the outputs from machine learning and artificial intelligence typically lack intuitive interpretation and unified representation. Such a data mining task is challenging due to the huge data nature of smart energy data.

Big data security and privacy: Although many security solutions have been proposed for SGs, they were not designed or customized specifically for energy big data. Attacks that make inferences directly from the energy big data can mislead the BDD so that fake data are unable to be detected. Also, the data can contain sensitive and private information of the customers and lead to usage pattern attacks. Most importantly, such data can be used to impact decision making on safe operation of the critical infrastructure.

Cyber-physical coupling modeling: One of the best known security features in SGs is tight cyber-physical coupling between the physical grid and cyber information, which exhibit multiple and distinct behavioral modalities and are deeply intertwined. A good understanding of it will be essential for ensuring the security of SG infrastructures.

Big Data Analytics and Applications in Smart Grid

Energy big data analytics is a very important research topic involving large distributed infrastructures, such as big data generation, transmission, storage, sharing, and processing. In addition to traditional challenges of big data analytics, energy big data analytics will also encounter difficulties in dealing with the unique features arising from tight cyber-physical coupling.

The required techniques involve a number of disciplines, including artificial intelligence, statistics, pattern recognition, machine learning, data mining, signal processing, and optimization and visualization methods. Big data analytics includes classification, aggregation, clustering, and data mining, as briefly described below.

One of the best known security features in SGs is a tight cyber-physical coupling between the physical grid and cyber information, which exhibit multiple and distinct behavioral modalities and are deeply intertwined. A good understanding of it will be essential for ensuring the security of SG infrastructures.

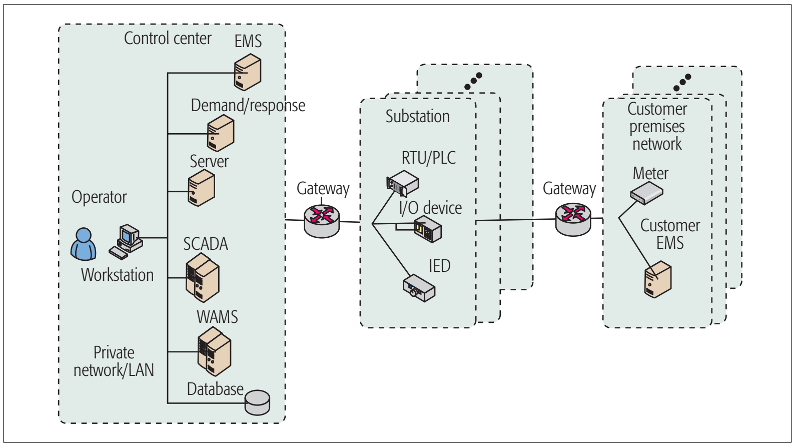


Figure 1. The management and control network of power grids with an emphasis on the distribution level.

Classification: Classification of a large volume of data is the process of organizing data according to its categories for its most effective and efficient use, also referred to as mining classification rules, a major application of data mining technology.

Aggregation: Data aggregation is a kind of data and information mining technique, where data is explored and presented in a report-based or shortened format to reduce computational cost.

Clustering: Clustering analysis can be used as an independent tool to obtain data distribution. Based on feature extraction and classification, the accuracy and efficiency of data mining can be improved.

Data mining: Via various methods, including artificial intelligence, machine learning, statistics, and database systems, useful patterns in large datasets can be extracted and transformed into a convenient and concise form.

To improve the reliability and efficiency of SG operation, power utilities are employing IT technology to develop big data applications [14].

We summarize some potential applications based on big data analytics in SG as follows:

- Load management with demand response
- Performance and efficiency analyses for power generation and storage systems
- Power grid optimization and capital expense minimization
- Large-scale and distributed state estimation based on AMI and smart devices
- Asset management by distributed islanding and aging transformer replacement
- Prediction and analysis of economic situation and social impact
- Pricing analysis and energy utilization
- Information provision for customers to better manage energy usage and bills, customer service enhancement, and customer behavior analysis
- Restoration spatial view of customer information, including trouble tickets, troubleshooting and fault localization, and real-time outage indication
- Scientific reasoning for policy making processes

Energy Big Data Attacks

System Model

The SG is a new electricity network, which encompasses advanced sensing and measurement technologies, ICTs, analytical and decision-making technologies, as well as the current power grid infrastructure. Figure 1 illustrates the management and control network of power grids with an emphasis on its distribution level. In the control center, the energy management system (EMS) consisting of BDD is a system of computer-aided tools used by operators to monitor, control, and optimize the performance of generation, transmission, and distribution of electrical power; the SCADA system is responsible for monitoring and control functions of the grid; wide area monitoring systems (WAMSs) employ new data acquisition technology based on phasor measurement and allow monitoring the conditions of a power system over a large-scale area to counteract grid abnormalities; and the database stores meter data, transmission admittance, topology information, system state, and so on. As a part of EMS applications, demand response provides an opportunity for consumers to play a role in the operation of the electric grid by reducing their electricity usage during peak load hours to save cost.

The programmable logic controllers (PLCs) and remote terminal units (RTUs) control devices autonomously without a master computer; the I/O devices are sensors and actuators; and the IEDs are microprocessor-based controllers of circuit breakers, feeders, substation transformers, capacitor banks, and phasor measurement units (PMUs). The EMS allows a customer to track its energy use in an easy format on computers or handheld devices.

Energy Big Data Replay Attack

The evolution from old power grids to SG brings new challenges in security. Hackers can eavesdrop or intercept metering data or steal big data from the distributed databases via malware. Normally, the grid parameters are unlikely to be known and often critically protected. Exposure of the structured data can cause losses in utilities or even a severe power imbalance problem. We demonstrate that a stealthy attack with both DC and AC state estimations can be successfully launched for misleading a power system through a replay mechanism. We call it an energy big data replay attack. The problem of interest can be formulated as follows.

Given a measurement vector set $\mathbf{z}_d$, $d = 1, 2, \ldots, D$, obtained from the energy big data, an energy big data attack can cheat the BDD as if no fabricated data exist. Or it can be detected by the BDD with a negligible probability. In addition to DC state estimation, the nonlinear AC state estimation is used inevitably in power systems because the AC state estimation has its advantages, including accuracy, ability against data manipulation attacks, and so on. Therefore, the attack should be able to pass the BDD using either DC or AC state estimation. For practicality, the power grid topology and transmission line admittances are not necessarily known to the attacker; therefore, this is a type of blind attack [11].

According to the criteria of a stealthy attack against AC state estimation, a perfect attack vector, **a**, should follow [6]

$$\mathbf{a} = \mathbf{h}(\mathbf{x}_a) - \mathbf{h}(\mathbf{x}), \tag{1}$$

where $\mathbf{h}(\cdot)$ denotes a general AC power flow model, and $\mathbf{x}_a$ and $\mathbf{x}$ denote the targeted and original state vectors of power systems, respectively.

The compromised measurement, $\mathbf{z}_a$, can be written as [6]

$$\mathbf{z}_a = \mathbf{z} + \mathbf{a} = \mathbf{h}(\mathbf{x}) + \mathbf{a} = \mathbf{h}(\mathbf{x}_a), \qquad (2)$$

where $\mathbf{z}$ denotes the original measurement vector. Figure 2 shows a geometric representation of the measurement vector $\mathbf{z}$, attack vector $\mathbf{a}$, and measurement vector under attack $\mathbf{z}_a$ in the AC power grid model between buses *i* and *j*. Notably, the AC power grid model is inherently nonlinear. For illustration purposes, the voltage amplitudes of two buses are normalized, the conductance and susceptance of the transmission line are 1.1350 and –4.7600, respectively, and a two-dimensional surface for the active power measurement vector $\mathbf{z}$ is assumed and presented. A similar two-dimensional surface for the reactive power measurement vector can also be demonstrated but omitted here. In view of the geometric representation, Eq. 2 indicates that the compromised measurement should lie on the surface of the AC power grid model, as shown in Fig. 2. Moreover, if a new compromised measurement is desirable and different from those in the observed data set, a tolerance mechanism can be introduced, provided that it is within a tolerable residue from the compromised measurement, which is typically related to a threshold of BDD.

Similarly, the criteria of a stealthy attack against the DC state estimation give the following relation [11]:

$$\mathbf{a} = \mathbf{Hc}, \qquad (3)$$

where $\mathbf{H}$ denotes the Jacobian matrix of the DC power flow model, and $\mathbf{c}$ is an arbitrary nonzero vector. Accordingly, substitute Eq. 3 into Eq. 2 and apply the DC model expression of $\mathbf{z}$. The compromised measurement in Eq. 2 also suggests that $\mathbf{z}_a$ should lie on the surface of the DC power grid model, which is inherently linear, as shown in Fig. 3. This is not surprising because the DC power flow model is a special case of the AC one.

Based on the aforementioned discussions, considering the measurement vector set $\mathbf{z}_d$, we propose an energy big data replay attack by formulating the attack vector as the difference between an observed measurement, $\mathbf{z}_d$, and the original measurement, vector $\mathbf{z}$. Here, the observed measurement $\mathbf{z}_d$ is treated as the compromised measurement $\mathbf{z}_a$. With the proposed attack vector, the compromised measurement will be positioned definitely on the surface of the power grid model. The selection of a specific $\mathbf{z}_d$ in the whole dataset can be done based on the maximum Euclidean distance between the compromised measurement and the original measurement vectors to impose a large abrupt change in the power system states. Or, on the contrary, the minimum distance rule can be adopted here to introduce a small change in the power system states and to reduce the possibility of being detected by an advanced detection mechanism.

Performance Evaluation

Monte Carlo simulations were conducted to assess the performance of the proposed big data attack (Big), random attack (Random), conventional DC attack (DC Conventional), and no attack (Ideal), which is used as a benchmark. The introduction of random, DC conventional, and Ideal attacks was already done in [11], and thus we do not repeat them here. The simulation results are evaluated in the IEEE 14-Bus electrical grid model. The measurements consist of active and reactive power flows at all branches. The number of simulations and the number of measurement vectors for each simulation run are 500 and 200, respectively. The impacts of measurement noise with zero-mean Gaussian distribution were evaluated.

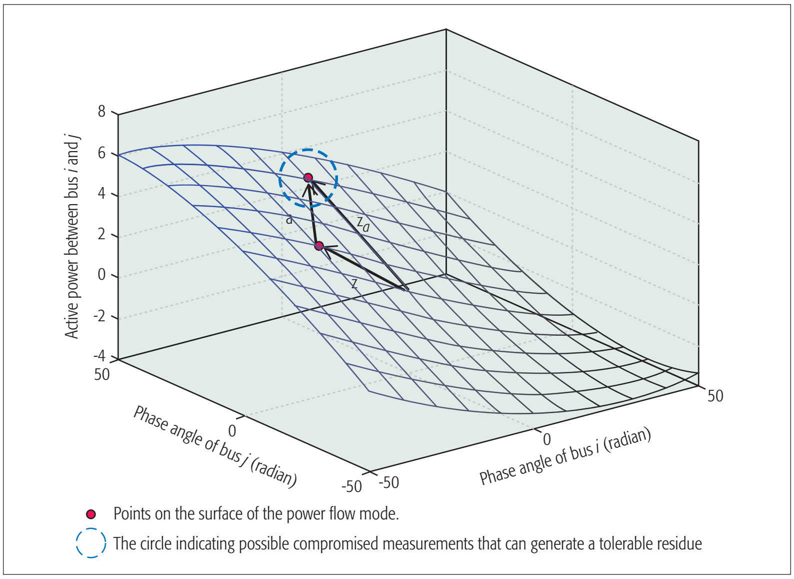


Figure 2. Geometric representation of the measurement vector $\mathbf{z}$, attack vector $\mathbf{a}$, and measurement vector under attack $\mathbf{z}_a$ in the AC power grid model.

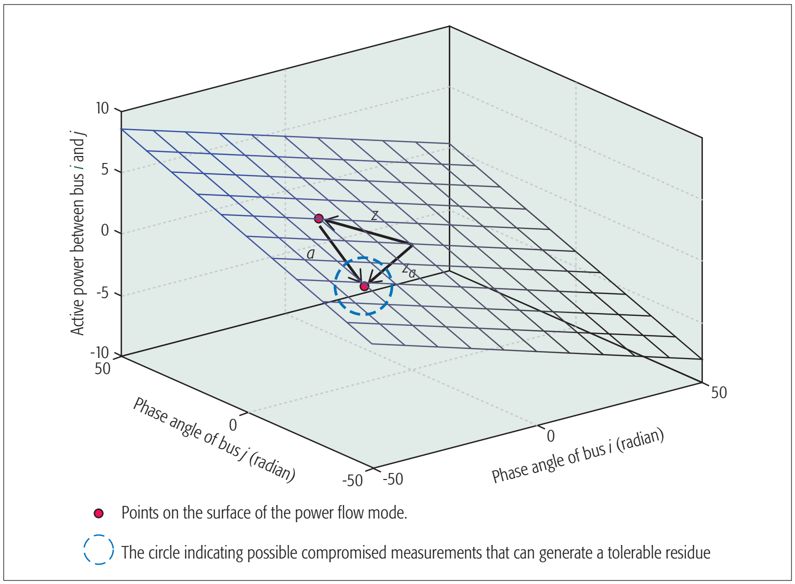


Figure 3. Geometric representation of the measurement vector $\mathbf{z}$, attack vector $\mathbf{a}$, and measurement vector under attack $\mathbf{z}_a$ in the DC power grid model.

Figures 4 and 5 plot the probability of missed detection, P_{miss}, vs. the decision threshold γ of BDD over the IEEE 14-Bus grid model against the DC and AC state estimations, respectively. The maximum distance rule for selecting the compromised measurement is adopted. As shown in Fig. 4, the random attack without taking the Jacobian matrix into consideration has the lowest P_{miss}; hence, it is not stealthy. The performance of the DC conventional attacks and that of the proposed big attacks coincide with that of the Ideal condition; therefore, they are indeed stealthy and perfect attacks. It is not surprising because the residue is ensured to be unaltered by the proposed scheme. To simplify the analysis, the proposed attack **a** satisfies Eq. 1. Then Eq. 2 guarantees that the compromised measurement lies on the surface of the power flow model so that the residue is unchanged. The proof follows. As shown in Fig. 5, the DC conventional and random attacks using a wrong power flow model have the lowest P_{miss}. The performance of the proposed big data attack is almost the same as that of the Ideal condition; therefore, it is still considered to be stealthy under the AC state estimation.

Therefore, the proposed algorithm is proved to be very flexible, requiring only measurement data, and applicable under DC or AC state estimations.

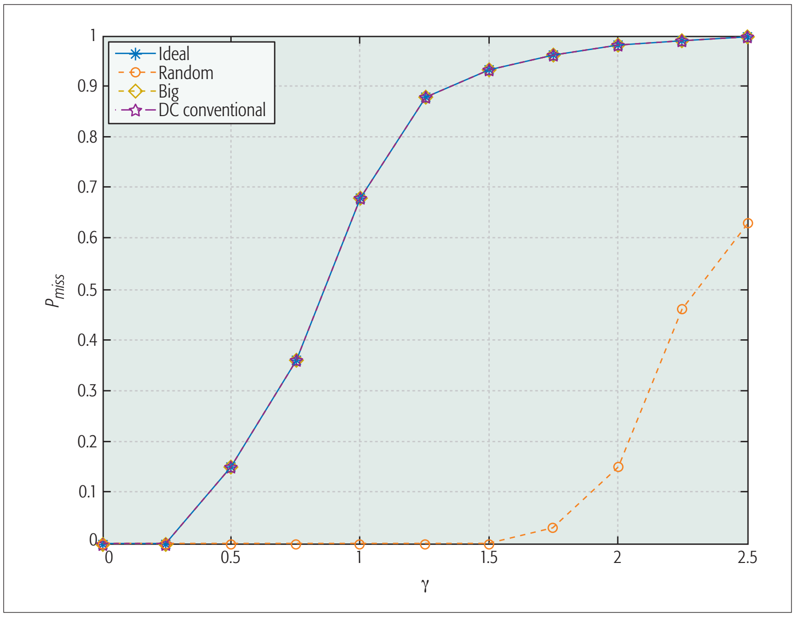


Figure 4. Probability of missed detection, P_{miss}, vs. decision threshold γ of BDD over the IEEE 14-Bus grid model against DC state estimation.

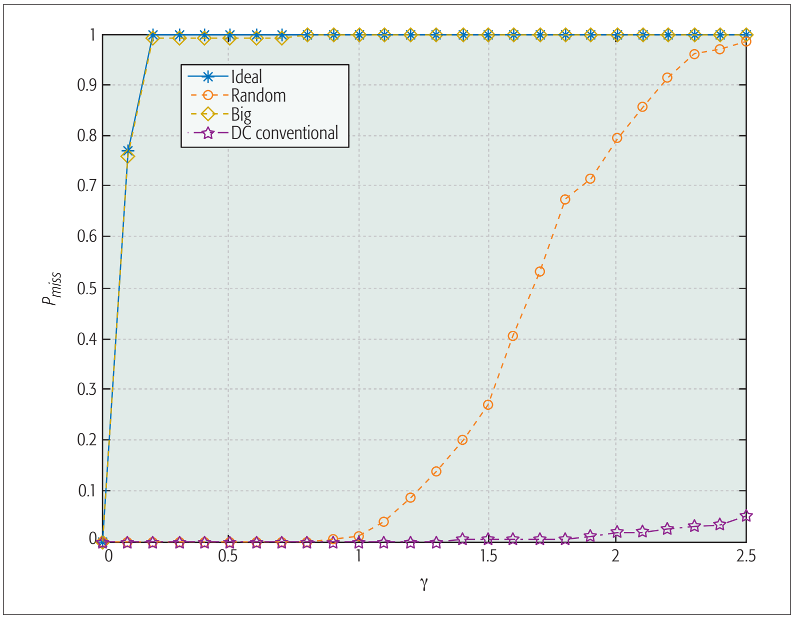


Figure 5. Probability of missed detection, P_{miss}, vs. decision threshold γ of BDD over the IEEE 14-Bus grid model against AC state estimation.

The future challenges of the big data attacks are outlined as follows:

- Sophisticated selection rules for the compromised measurement that can significantly confuse a power system need to be investigated further. For example, a random selection approach can be one of them.
- New powerful metrics might exist in addition to the proposed one based on the Euclidean distance.
- The proposed attack opens a new research direction from the attackers' viewpoints. New defense mechanisms are required to deal with it efficiently.

Future Research Directions in IoT-Based Smart Grid

In addition to removal of business and political barriers, governmental efforts should pursue several goals concurrently, including regulations, universal standards, failure recovery mechanisms, and so on. Several challenges can be identified as follows [9].

Communication technologies: The success of IoT-based SG depends strongly on uninterrupted communications of its connected devices. A huge amount of energy big data related to monitoring and control will be transmitted using wireless and wireline communication infrastructures, such as Wi-Fi, Bluetooth, ZigBee, cellular, WiMAX, PLC, and fiber optics. Cognitive radio (CR) networking was recognized as a prominent technology to address communication requirements of IoT-based SGs [15].

Heterogeneity: Due to the discrepancy on the resources that the devices and communication technologies use in the SG, achieving end-to-end security and connectivity is a challenging task, requiring a complex cyber-physical coupling model. Co-design of energy big data analytics and security mechanism can minimize security risk. Moreover, regional differences in electric grid topologies require diverse technologies to resolve interconnection issues.

Scalability: Independent random events can aggregate to yield large-scale catastrophic failures in the grid and trigger cascading events. Particularly, scalable key management, authentication [10], and privacy solutions are required for the large-scale deployment of SG.

Constrained resources: SG devices are resource constrained. Security solutions, such as authentication, for a large number of nodes in SG have become a challenging issue [10].

Interoperability: Legacy systems were deployed based on proprietary hardware and software. The implementation of IoT-based SG should also be coordinated with governmental efforts under national energy policies, national security, economic growth, and energy independence. As a result, they pose unique challenges to create a suite of standards for the interoperability and backward compatibility in SG.

Trust management: Trust must be established across different SG domains and/or levels, including different utilities and electricity generation chains. Building the trust between different domains is a challenge, especially in a large-scale IoT network with a large number of low-end SG devices.

Latency constraint: Essential information should be stored, processed, and extracted in a timely manner. Therefore, modern big data analytics is an important method for the intelligence and decision making in the SG.

Service on demand: Cloud computing architecture provides shared processing resources and data for energy big data analytics, as shown in Fig. 6. A new platform is needed to deal with big data and security concerns in a prompt fashion. The cloud control center can provide different levels of service, such as infrastructure as a service, platform as a service, and software as a service for traditional utility and local control centers, and even customers. The third party services may include a weather forecast and authentication center with the key generator. Based on historical data and information from the third party service, big data analytics are applied at the cloud control center for energy forecast, security analysis, and so on. The local control centers are distributed for better scalability and reliability. If a local control center is unavailable due to maintenance, attacks, or natural disasters, other local control centers can take over the control.

Network-based threats detection: We have shown a new big data attack in this work. Additional attacks can also appear. Besides, we need to rely on automated detection schemes to respond to network-based threats. The vulnerabilities of grids should be detected early enough. Quick and auto-recovery mechanisms need further research efforts. Furthermore, the mindset of utilities is still focused on reliability under natural disasters instead of security threats from adversaries. Also, very few studies have been carried out on key management schemes for AMI and wide area measurement network entities. Besides, a distributed security solution is needed to protect essential/privacy information.

Self-healing protection systems: Relay applications for the protection of power systems have been used for over 100 years. Advanced algorithms, such as islanding protection employing IEDs and PMUs with sensors, are important for SG.

Conclusions

The SG can benefit from the IoT technology, where smart devices are integrated with pervasive connectivity. Security is the main concern for the IoT-based SG, which works in a complex cyber-physical model. In this article, we have reviewed the main security issues and challenges for the IoT-based SGs, and discussed the problems with energy big data analytics. While enjoying the benefits of SG, we have to prevent individual privacy intrusion and keep the data from being abused. In particular, we have demonstrated a big data attack that can be launched by knowing only limited information. The work presented in this article can raise awareness of the security concerns in the IoT-based SG.

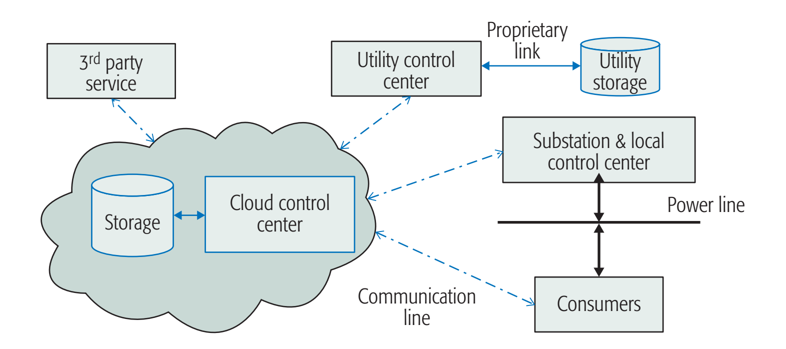


Figure 6. A cloud computing platform for smart grid applications with an emphasis on the distribution level.

References

[1] Y. Yan *et al.*, "A Survey on Smart Grid Communication Infrastructures: Motivations, Requirements and Challenges," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 1, 1st qtr. 2013, pp. 5–20.

[2] F. Ghavimi and H. H. Chen, "M2M Communications in 3GPP LTE/LTE-A Networks: Architectures, Service Requirements, Challenges, and Applications," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 2, 2nd qtr. 2015, pp. 525–49.

[3] C. Lai *et al.*, "Toward secure Large-Scale Machine-to-Machine Communications in 3GPP Networks: Challenges and Solutions," *IEEE Commun. Mag.*, vol. 53, no. 12, Dec. 2015, pp. 12–19.

[4] Y. Yan *et al.*, "A Survey on Cyber Security for Smart Grid Communications," *IEEE Commun. Surveys & Tutorials*, vol. 14, no. 4, 4th qtr. 2012, pp. 998–1010.

[5] N. Li *et al.*, "Researches on Data Processing and Data Preventing Technologies in the Environment of Big Data in Power System," *Proc. DRPT '15*, Nov. 2015, pp. 2491–94.

[6] Md. A. Rahman and H. Mohsenian-Rad, "False Data Injection Attacks Against Nonlinear State Estimation in Smart Power Grids," *Proc. IEEE PES General Meeting '13*, July 2013, pp. 1–5.

[7] X. Li *et al.*, "Securing Smart Grid: Cyber Attacks, Countermeasures, and Challenges," *IEEE Commun. Mag.*, vol. 50, no. 8, Aug. 2012, pp. 38–45.

[8] R. Ma *et al.*, "Smart Grid Communication: Its Challenges and Opportunities," *IEEE Trans. Smart Grid*, vol. 4, no. 1, Mar. 2013, pp. 36-46.

[9] C. Bekara, "Security Issues and Challenges for the IoT-Based Smart Grid," *Elsevier Procedia Computer Science*, vol. 34, 2014, pp. 532–37.

[10] W. L. Chin, Y. H. Lin, and H. H. Chen, "A Framework of Machine-to-Machine Authentication in Smart Grid: A Two-Layer Approach," *IEEE Commun. Mag.*, vol. 54, no. 12, Dec. 2016, pp. 102–07.

[11] Z. H. Yu and W. L. Chin, "Blind False Data Injection Attack Using PCA Approximation Method in Smart Grid," *IEEE Trans. Smart Grid*, vol. 6, no. 3, May 2015, pp. 1219–26.

[12] K. Wang *et al.*, "Strategic Honeypot Game Model for Distributed Denial of Service Attacks in the Smart grid," *IEEE Trans. Smart Grid*, DOI: 10.1109/TSG.2017.2670144, 2017.

[13] H. Jiang *et al.*, "Energy Big Data: A Survey," *IEEE Access*, vol. 4, 2016, pp. 3844–61.

[14] C. S. Lai and L. L. Lai, "Application of Big Data in Smart Grid," *Proc. IEEE SMC '15*, Hong Kong, China, Sept. 2015, pp. 665–70.

[15] T. N. Le, W. L. Chin, and H. H. Chen, "Standardization and Security for Smart Grid Communications Based on Cognitive Radio Technologies – A Comprehensive Survey," *IEEE Commun. Surveys & Tutorials*, vol. 19, no. 1, 2st qtr. 2017, pp. 423–45.

Biographies

Wen-Long Chin (wlchin@mail.ncku.edu.tw) received his M.S. degree in electrical engineering from National Taiwan University and his Ph.D. degree in electronics engineering from National Chiao Tung University in 1996 and 2008, respectively. He is now an associate professor in the Department of Engineering Science, National Cheng Kung University. Before holding this faculty position, he worked at Hsinchu Science Park, Taiwan. He serves as an Associate Editor of *IEEE Access*.

Wan Li (kghs980824@yahoo.com.tw) received her B.Sc. degree in electrical engineering from National University of Tainan, Taiwan, in 2016. Now she is a first-year graduate student in engineering science at National Cheng Kung University, Tainan City, Taiwan. Her research interests include the Internet of Things and smart grid.

Hsiao-Hwa Chen [S'89, M'91, SM'00, F'10] (hshwchen@ieee.org, hshwchen@mail.ncku.edu.tw) is currently a Distinguished Professor in the Department of Engineering Science, National Cheng Kung University. He is the founding Editor-in-Chief of Wiley's *Security and Communication Networks Journal*. He was the recipient of the 2016 IEEE Jack Neubauer Memorial Award. He served as Editor-in-Chief of *IEEE Wireless Communications* from 2012 to 2015, and as an elected Member at Large of IEEE ComSoc from 2013 to 2016. He is a Fellow of IET.

Defense Mechanisms against Data Injection Attacks in Smart Grid Networks

Jing Jiang and Yi Qian

Focusing on signal processing techniques, the authors introduce an adaptive scheme on detection of injected bad data at the control center. This scheme takes the power measurements of two sequential data collection slots into account, and detects data injection attacks by monitoring the measurement variations and state changes between the two time slots.

Abstract

In the smart grid, bidirectional information exchange among customers, operators, and control devices significantly improves the efficiency of energy supplying and consumption. However, integration of intelligence and cyber systems into a power grid can lead to serious cyber security challenges and makes the overall system more vulnerable to cyber attacks. To address this challenging issue, this article presents defense mechanisms to either protect the system from attackers in advance or detect the existence of data injection attacks to improve the smart grid security. Focusing on signal processing techniques, this article introduces an adaptive scheme on detection of injected bad data at the control center. This scheme takes the power measurements of two sequential data collection slots into account, and detects data injection attacks by monitoring the measurement variations and state changes between the two time slots. The proposed scheme has the capability of adaptively detecting attacks including both non-stealthy attacks and stealthy attacks. Stealthy attacks are proved impossible to detect using conventional residual-based methods, and can cause more dangerous effects on power systems than non-stealthy attacks. It is demonstrated that the proposed scheme can also be used for attack classification to help system operators prioritize their actions to better protect their systems, and is therefore very valuable in practical smart grid systems.

Introduction

The smart grid is a modernized power grid that uses information and communications technology to gather and act on information for improving the efficiency, robustness, economics, and sustainability of energy distribution and management [1]. The bidirectional information exchange among customers, operators, and control devices offers a more efficient way of energy supplying and consumption: On the operator side, equipment can be intelligently managed, and energy supplying flexibility can be significantly improved. On the consumer side, both the user experience and billing system can be enhanced [2]. The data generated in a smart grid is much more than that generated in a traditional power grid due to this continuous bidirectional information exchange [3, 4]. The Internet of Things (IoT) enables the transfer of such high volume data, and makes the grid infrastructure, meters, substations, and buildings virtually interconnected through the Internet or peer-to-peer connections [5]. IoT can be a valuable solution to support the development of smart grid.

However, by integrating a physical system (power grid) with a cyber system (IoT), a smart grid presents significant cyber security challenges and makes the overall system more vulnerable to cyber attacks. For instance, in December 2015, a cyber attack of a power system was reported in Ukraine, which caused a power cut lasting several hours and affecting 80,000 customers. During the attack, 103 cities were completely blacked out, and the affected control centers were not fully operational even two months later. In addition, according to data provided by the United States Computer Emergency Readiness Team, there were 79 cyber hacking incidents targeted at the energy sector in 2014 [6]. Such attacks could maliciously manipulate the electricity price in the power market, or even cause a regional blackout (taking Ukraine as an example), and result in serious social and economic consequences. Thus, IoT-enabled smart grids must incorporate appropriate cyber protection mechanisms for detecting and identifying such malicious data attacks to improve smart grid security.

To maintain normal operations of the smart grid, the power systems are continuously monitored and controlled by supervisory control and data acquisition (SCADA) systems and energy management systems (EMSs) [7]. In particular, the SCADA host receives real-time measurements (typically transmission line power flows and bus line power flows) from remote meters or sensors. These measurements are then processed at the state estimator for estimating the system states and building real-time electricity network models [8]. These state estimates are crucial, and must be passed to enable EMS application functions, such as automatic generation control and optimal power flow, to control the physical aspects of power grids. We consider a smart grid comprising the power system, communication network, and control center. An attacker may launch attacks by hacking a few meters or sensors to distort the measurements. Moreover, the communication links are also vulnerable to data injection attacks where measurements may be altered during data transmission [9]. Bad data injection attacks can result in the state estimator producing incorrect system state estimates, leading to poor control

Digital Object Identifier: 10.1109/MCOM.2017.1700180

Jing Jiang is with Durham University; Yi Qian is with the University of Nebraska-Lincoln.

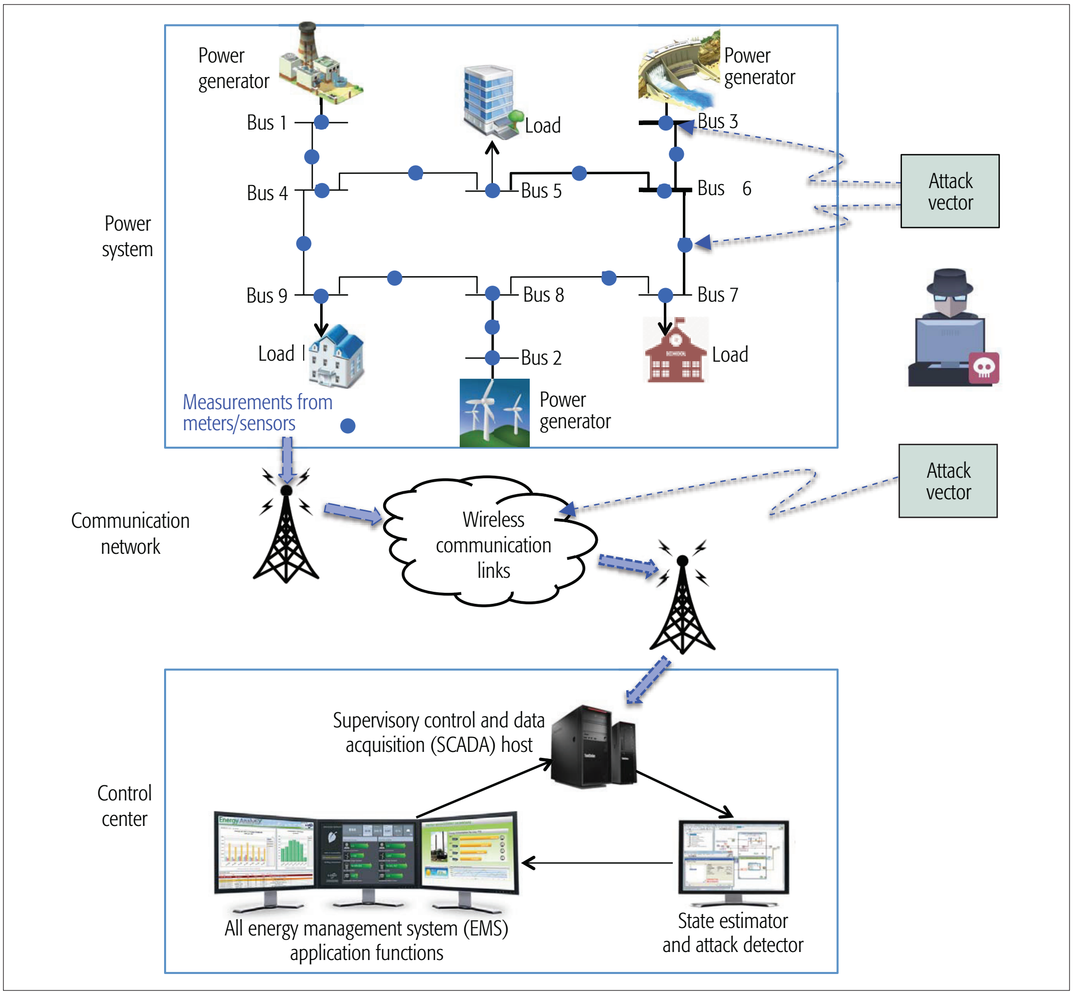

Power generator
Bus 1
Load
Power generator
Bus 3
Bus 4
Bus 5
Bus 6
Attack vector
Power system
Bus 9
Bus 8
Bus 7
Load
Bus 2
Load
Measurements from meters/sensors
Power generator
Communication network
Wireless communication links
Attack vector
Control center
Supervisory control and data acquisition (SCADA) host
All energy management system (EMS) application functions
State estimator and attack detector

Figure 1. A block architecture diagram of the power system, communication network, and control center of a smart grid. An IEEE 9-bus system is chosen to illustrate the power system.

The bidirectional information exchange among customers, operators, and control devices offers a more efficient way of energy supplying and consumption: On the operator side, equipment can be intelligently managed, and energy supplying flexibility can be significantly improved. On the consumer side, both the user experience and billing system can be enhanced.

decisions or a major malfunction or even blackout. Other, non-malicious, events can also result in a bad data injection. For example, a tree falling on a transmission line will cause a sudden and large change in some measurements. Such an event is referred to as an accident, which is also a type of data injection. It is desirable to protect power systems from data injection attacks in advance or detect bad data during the state estimation process at the control center.

In this article, we focus on data injection attacks and defense mechanisms in smart grid networks. Some preliminary works that include the problem formulation of state estimation and types of attacks are studied. Depending on whether the power grid topology information is known by attackers or not, data injection attacks are divided into two types, stealthy attacks and non-stealthy attacks. We then investigate defense mechanisms to protect power systems from these attacks. Traditionally, stealthy attacks are impossible to detect using conventional residual-based methods. We thus introduce a novel scheme to adaptively detect and classify data injection attacks including both non-stealthy and stealthy attacks. This scheme takes the measurements of two sequential data collection slots into account, and detects data injection attacks by monitoring the measurement variations and state changes between the two time slots. Using this scheme, once the attack type is identified, system operators can prioritize their actions or resources to better protect their systems.

The rest of this article is organized as follows. We first present the system architecture and introduce state estimation and data attacks in the smart grid. Two categories of defense mechanisms are then discussed. After that, we propose an adaptive scheme for detection of data injection attacks, and demonstrate the benefits of this scheme compared to conventional methods. Finally, this article is concluded.

State Estimation and Data Injection Attacks

In this section, we illustrate the system architecture and introduce the problem formulation of state estimation. Two types of data injection attacks, non-stealthy and stealthy attacks, are then defined.

System Architecture

Figure 1 shows a block architecture diagram of a power system, communication network, and control center of a smart grid. In order to clearly demonstrate the power system, a small-scale IEEE 9-bus system is employed that consists of three different types of power generators and three various loads. The power system is monitored by a control system, which comprises a SCADA host and a remote sensing system providing power measurement data to the SCADA host via a communication network. The remote sensing system comprises a plurality of remote sensors or meters. As shown in Fig. 1, the remote sensors come in two varieties: transmission line flow sensors, which measure the power flow through a single transmission line, and bus injection sensors, which measure the power injection flow from all transmission lines connected to a single bus. At one data collection slot, the measurements from these sensors will be transmitted through a communication network. A wireless communication network is considered in Fig. 1, as it can offer widespread access, great flexibility, and quick deployment. WiMAX or cellular network communications (e.g., third or fourth generation, 3G or 4G) can provide the wireless communication solutions [2]. In the control center, the real-time power measurements received by the SCADA host are then processed at the state estimator to estimate the system states and build real-time electricity network models. As shown in Fig. 1, an attacker may launch a data injection attack by hacking a few sensors to distort the measurements, or alter measurements during their transmission in the communication links.

State Estimation

At the control center, operators need to know the voltage phase angles of all buses to make control and operation decisions. However, it is difficult for sensors to directly measure phase angles [10]. The control center thus uses a state estimation technique to estimate the system states (typically voltage phase angles) through processing the set of real-time power measurements received from sensors.

At the control center, operators need to know the voltage phase angles of all buses to make control and operation decisions. However, it is difficult for sensors to directly measure phase angles. The control center thus uses a state estimation technique to estimate the system states through processing the set of real-time power measurements received from sensors.

In power systems, the transmitted power from one bus to another depends on the voltage amplitudes and voltage phase differences between the two buses, and also relates to the reactance of the transmission line between these buses. In power flow analysis, it is usually considered that the voltage phase differences are relatively small and the voltage amplitudes are normalized to unit [8], such that a linear relation exists between the power measurement and the voltage phase difference. We thus apply a linearized power flow model, which is widely used for real-time analysis of state estimation in power systems [11].

Using the linearized power flow model, the received power measurements at the SCADA host can be represented in a vector-matrix form as $\mathbf{z} = \mathbf{Hx} + \mathbf{u}$, where $\mathbf{z}$ denotes the measurement vector, which includes power flow measurements on transmission lines and power injection measurements at buses. The system state vector is represented by $\mathbf{x}$, and the vector $\mathbf{u}$ represents the Gaussian noise with a zero mean and a covariance matrix $\mathbf{u}$. The matrix $\mathbf{H}$ is the measurement matrix, which is assumed to be fully known to the system operators; attackers may or may not know this measurement matrix [11]. In the power system, the network connectivity can be described by an oriented incidence matrix $\mathbf{M}$; each column of $\mathbf{M}$ corresponds to one power transmission line, and the number of rows represents the number of buses. The physical properties of the transmission lines can be described by a nonsingular diagonal matrix $\mathbf{N}$, of which diagonal entries equal admittances of the transmission lines. The matrix $\mathbf{H}$ can be constructed by $\mathbf{H} = [\mathbf{NM}^T, \mathbf{MNM}^T]^T$ [9]. That is, power flow measurements on transmission lines are obtained from $\mathbf{MNM}^T\mathbf{x}$, and power injection measurements at buses can be computed from $\mathbf{NM}^T\mathbf{x}$.

State estimation uses the received measurements $\mathbf{z}$ to timely estimate the power system states $\mathbf{x}$. The vector $\mathbf{x}$ can be computed using the weighted least square method: $\hat{\mathbf{x}} = (\mathbf{H}^T\mathbf{UH})^{-1}\mathbf{H}^T\mathbf{Uz}$.

Non-stealthy Attacks and Stealthy Attacks

If a bad data injection attack occurs, the received power measurement vector will include a bad data vector maliciously injected by the attacker. That is, $\mathbf{z} = \mathbf{Hx} + \mathbf{a} + \mathbf{u}$, where $\mathbf{a}$ denotes the bad data vector. Conventional methods to detect bad data injection are mostly based on residual tests. Residual refers to the difference between the measurement vector $\mathbf{z}$ and the calculated value from the estimated state (i.e., $\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}$). The largest normalized residual test can be used to detect bad data injection to see if the largest absolute value of the elements in normalized residual is greater than a pre-defined threshold. If the largest normalized residual is larger than the threshold, the corresponding measurement will be considered as bad data and reported to system operators.

Depending on whether or not the bad data attacks are detectable by conventional residual tests, we define the following two types of attacks.

Non-Stealthy Data Injection Attacks: These are defined as attacks detectable by conventional residual-test methods [8]. In this case, the measurement matrix $\mathbf{H}$ is not known to the attackers. The attackers simply generate random attack vectors and manipulate the meter readings.

Stealthy Data Injection Attacks: These are defined as attacks not detectable by conventional bad data detection methods. In this case, attackers are assumed to be familiar with the power grid topology information or know the measurement matrix. They can carefully design the malicious data and let $\mathbf{a} = \mathbf{Hc}$, where $\mathbf{c} \in \mathbb{R}^n$ can be any arbitrary vector [12]. The measurement vector can then be written as $\mathbf{z} = \mathbf{H}(\mathbf{x} + \mathbf{c}) + \mathbf{u}$. Such attacks can bypass the conventional residual-test detection methods, and the control center would believe that the true state is $(\mathbf{x} + \mathbf{c})$.

Defense Mechanisms

Since bad data injection attacks can result in poor control decisions or a major malfunction or even blackout, it is crucially important to have appropriate defense mechanisms to either protect the system from attackers in advance or identify the existence of bad data injection attacks during the state estimation process [9]. Defense mechanisms can be divided into two categories. One is to deploy advanced measurement units, such as phasor measurement units (PMUs), at various locations to protect the system from attackers in advance. The other is to adopt advanced signal processing techniques at the control center to identify bad data injection attacks.

Deployment of Advanced Measurement Units

The mechanisms of deploying advanced measurement units, such as PMUs, are introduced in [13–15]. PMUs measure voltages and currents on a power grid using a common time source based on global positioning system (GPS) time, and thus have the capability of providing accurately time-stamped measurements for geographically dispersed nodes. Consequently, PMUs are typically robust against data injection attacks and have the measurements secured. In practice, PMUs are very expensive; it is not feasible to deploy enough PMUs to secure all measurements in a grid network. It is demonstrated in [14] that it is possible to defend against malicious data injection by either protecting a subset of existing measurements or placing additional secure PMUs on a fraction of buses. The challenge, however, is that selecting such subsets is a high-complexity problem, and recent studies have proposed several methods on how to address this issue. For instance, [14] proposes a fast greedy algorithm to select a subset of measurements to be protected, [13] uses graphical characterization to study defending mechanisms with a minimum number of secure measurements, and [15] provides a semidefinite programming optimal PMU layout approach considering the impact of restricted channel limits. Due to the high cost, the approach of deploying advanced measurement units to defense data injections will be more suitable for power systems that have great social and economic impacts, but for a general power system, it will be restricted by limited budget.

Due to the high cost, the approach of deploying advanced measurement units to defense data injections will be more suitable for power systems that have great social and economic impacts, but for a general power system, it will be restricted by limited budget.

Adoption of Signal Processing Techniques

The mechanisms investigating advanced signal processing techniques are to detect the injected bad data at the control center and discard these data from measurements. The attack detector (as shown in Fig. 1) needs to reliably detect a data injection attack in the event of an attack. Either an attack event has occurred or it has not. The detector either identifies an attack or does not. There are four possible outcomes: hit (the attack presents and the detector identifies), miss (the attack presents and the detector fails to identify), false alarm (there is no attack and the detector wrongly identifies one), and correct rejection (there is no attack and the detector identifies no attack). We define probability of detection to indicate the first case and probability of false alarm to indicate the probability of the third case. The two probabilities can be used as indicators to compare the performance of different detection methods.

As mentioned earlier, conventional methods to detect bad data are mostly based on residual tests. When a stealthy data injection attack happens, the residual would not change compared to the no-attack case, and the system would not report any abnormal state. Besides conventional methods, some other advanced signal processing techniques are considered to improve detection accuracy. In [12], machine learning (ML) is proposed for detecting stealthy attacks. This ML technique relies on a set of historical data that is used for learning and validating data to detect the attacks in new measurements, and the learning efficiency needs to be improved. A cumulative-sum-based (CS) approach is proposed in [8] aiming to minimize the detection time subject to certain detection error constraints; but this CS approach focuses on non-stealthy attacks. Exploiting the low rank structure of temporal erroneous-free measurements and sparsity of malicious attacks, defense mechanisms are proposed in [9, 10], where methods of constructing sparse stealthy attacks are also studied in [9]; these mechanisms have a strong assumption that the attack matrix must be spare, which is not robust against attackers with strong capability of launching cyber attacks. One novel mechanism to adaptively detect and classify data injection attacks (including non-stealthy attacks and stealthy attacks) is presented next.

An Adaptive Scheme for Detection of Data Injections

In this section, an adaptive scheme for detection of data injections is presented. It takes the measurements of two sequential data collection slots into account, and the equation of received measurements can be written as state-space equations with discrete time index i, as $\boldsymbol{z}_i = \boldsymbol{H}\boldsymbol{x}_i + \boldsymbol{u}_i$ and $\boldsymbol{x}_i = \boldsymbol{x}_{i-1} + \Delta\boldsymbol{x}_i$, where $\Delta\boldsymbol{x}_i$ is the state change vector representing the system state changes from the last data collection slot $i - 1$ to the current data collection slot i. Current smart meters support 15-minute-interval data collection frequency, and the frequency is likely to improve further for achieving advanced smart grid functionalities. Compared to the values of system state, the values of system state changes are relatively small, that is, the system state generally varies in a small dynamic range. The state change vector $\Delta\boldsymbol{x}_i$ follows a certain distribution and is here initialized to be normal distribution with zero mean. In addition, as the measurement matrix $\boldsymbol{H}$ is related to the power network connectivity and physical properties of the transmission lines, $\boldsymbol{H}$ generally remains unchanged for the two small sequential data collection slots. Any updates of the measurement matrix will be reported to the control center, and the updated one will be used for state estimation and attack detection.

We monitor measurement change and residual and state change between two data collection slots to detect and classify non-stealthy and stealthy data injection attacks. Let $\boldsymbol{w}_i$ present the measurement change vector, which is the difference between the current power flow measurement vector and the calculated value of the last estimated state (i.e., $\boldsymbol{w}_i = \boldsymbol{z}_i - \boldsymbol{H}\hat{\boldsymbol{x}}_{i-1}$). We can compute the last estimated state using the weighted least square method shown earlier, and obtain $\boldsymbol{w}_i = \boldsymbol{H}\Delta\boldsymbol{x}_i + (\boldsymbol{u}_i - \boldsymbol{u}_{i-1})$. As $\boldsymbol{u}_i$ and $\boldsymbol{u}_{i-1}$ are independent Gaussian noise vectors at the two sequential data collection slots, $(\boldsymbol{u}_i - \boldsymbol{u}_{i-1})$ is also Gaussian distributed with a zero mean and a covariance matrix $2\boldsymbol{U}$. The state change vector can then be estimated from the measurement change vector $\boldsymbol{w}_i$ by using the weighted least square method. Furthermore, we define the measurement change residual vector $\boldsymbol{r}_i = \boldsymbol{w}_i - \boldsymbol{H}\Delta\hat{\boldsymbol{x}}_i$, and compute its Euclidean norm to detect the presence of non-stealthy data injection attack. That is, if the Euclidean norm of $\boldsymbol{r}_i$ is greater than a pre-defined threshold τ_1, the presence of non-stealthy data injection will be inferred and reported to system operators. Note

```
I. Initialization: $i = 0$, collect historical estimated state vector.
II. Repeat the procedure for detecting bad data injection attacks:
  1. Update $i \leftarrow i + 1$;
  2. Obtain current measurement vector $\boldsymbol{z}_i$;
  3. Calculate the measurement change vector $\boldsymbol{w}_i$;
  4. Estimate the state change $\Delta\hat{\boldsymbol{x}}_i$ using the weighted least-square method;
  5. Calculate the measurement change residual $\boldsymbol{r}_i$;
  6. Identify and classify data attacks:
       if the Euclidean norm of $\boldsymbol{r}_i$ is larger than $\tau_1$ then
         report non-stealthy data injection attack and exit
       else if the Euclidean norm of $\Delta\hat{\boldsymbol{x}}_i$ is larger than $\tau_2$ then
         report stealthy data injection attack and exit
       else make the decision of no attack, obtain the estimated state $\hat{\boldsymbol{x}}_i$; update the last-time-slot
       estimated state $\hat{\boldsymbol{x}}_{i-1} = \hat{\boldsymbol{x}}_i$, and continue the state estimation process
III. Until non-stealthy or stealthy data injection attack is determined.
IV. Terminate the attack detection process and report the detected data injection attack to
system operators.
V. System operators distinguish between data injection attacks and electrical accidents.
```

Algorithm 1. Adaptive detection of data injection attacks.

that the Euclidean norm is also called ℓ^2 distance or ℓ^2 norm. Besides the Euclidean norm test, other test methods on the residual (e.g., the largest normalized residual test) can also be used to detect the presence of non-stealthy data injection. The selection of the threshold τ_1 is based on history and trade-off between the probability of detection and probability of false alarm. In addition, if no data injection is inferred (i.e., the Euclidean norm of $\boldsymbol{r}_i$ is equal to or less than τ_1), we use the Euclidean norm of state change $\Delta\hat{\boldsymbol{x}}_i$ to detect the presence of stealthy data injection attack. If the Euclidean norm of $\Delta\hat{\boldsymbol{x}}_i$ is greater than a pre-defined threshold τ_2, the presence of stealthy data injection will be inferred and reported to system operators.

When attackers launch stealthy data injection, we have the state change vector as $\Delta\boldsymbol{x}_i + \boldsymbol{c}_i$, compared to $\Delta\boldsymbol{x}_i$ for the non-attack case. Due to the fact that the vector $\Delta\boldsymbol{x}_i + \boldsymbol{c}_i$ does not exhibit the same distribution feature as the vector $\Delta\boldsymbol{x}_i$, using detection algorithms, the existence of $\boldsymbol{c}_i$ can be detected at a certain successful probability. The detection probability of the proposed scheme will be higher if the elements of $\boldsymbol{c}_i$ are larger. A stealthy attack with larger $\boldsymbol{c}_i$ can cause greater system state change, and thus is more dangerous to the power system operations.

Referring to Algorithm 1, the steps of the proposed scheme are as follows. We initialize time index $i = 0$ and collect historical estimated state vector $\hat{\boldsymbol{x}}_0$. Then the procedure of detection for bad data attacks or electrical accidents is carried out: The time index is updated as $i = i + 1$, and the current measurement vector $\boldsymbol{z}_i$ sent by remote sensors is obtained. We calculate the measurement change vector $\boldsymbol{w}_i$, which is the difference between the vector $\boldsymbol{z}_i$ and $\boldsymbol{H}\hat{\boldsymbol{x}}_{i-1}$. The state change vector $\Delta\hat{\boldsymbol{x}}_i$ is estimated from $\boldsymbol{w}_i$ by using the weighted least square method. We then calculate the measurement change residual vector $\boldsymbol{r}_i$, which is the difference between the vector $\boldsymbol{w}_i$ and $\boldsymbol{H}\Delta\hat{\boldsymbol{x}}_i$. As shown in step II-6 of Algorithm 1, data attacks are then identified based on the estimated state change vector $\Delta\hat{\boldsymbol{x}}_i$ and the measurement change residual vector $\boldsymbol{r}_i$ by using the Euclidean norm method. Here the largest normalized method (which compares the largest absolute value of elements in a vector with a threshold) can also be used. Data attacks will be classified into non-stealthy attack, stealthy attack, and no attack. If no attack is identified, the process of state estimation will continue. If non-stealthy or stealthy attack is determined, the process of attack detection will be terminated and the detected data injection attack will be reported to system operators. In step IV, for distinguishing between data injection attacks and electrical accidents, system operators can either send staff to verify or wait for receiving reports from secure devices, such as from PMUs or intelligent electronic devices.

The proposed scheme differs from existing relevant methods in three aspects. First, different from related works that only process measurements collected at one single time slot, this scheme takes the measurements of two sequential data collection slots into account, and detects injection attacks by monitoring the measurement variations and state changes between the two slots. Next, the proposed scheme can self-adaptively detect both non-stealthy and stealthy injection attacks; the latter were proved impossible to detect using conventional methods. Furthermore, the proposed scheme can identify the type of data attacks. Since stealthy attacks are more dangerous to power system operations than non-stealthy ones, it is crucial for operators to know the attack type and then prioritize their actions or resources to better protect their systems and reduce the chance of future stealthy attacks.

Performance Evaluation

In this section, we evaluate the performance of a conventional detection method and the proposed scheme for detecting both non-stealthy and stealthy attacks based on IEEE test systems. The MATLAB package MATPOWER is used to simulate the operation of the power system. The signal-to-noise ratio (SNR) considered in the simulations indicates the power level of true measurements to the power level of noise. For bad data injection attacks, both non-stealthy and stealthy attacks of various attack severity levels are considered. The attack-to-noise ratio (ANR) is used to indicate the attack severity level, defined as the ratio of attack power level to the noise power level. We use receiver operating characteristic (ROC) curves to illustrate the performance of a detector as the discrimination threshold is varied. The curve is generated by plotting the probability of detection against the probability of false alarm at various threshold settings. The probability of detection indicates the probability of saying that an "attack" is present given that an "attack" event actually occurred. The probability of false alarm is the probability of saying that an "attack" is present given that a "no attack" event actually occurred.

Referring again to the IEEE 9-bus test system and power system shown in Fig. 1, there are 9 transmission lines and thus 18 measurement elements in total for one data collection slot. Figure 2a shows the ROC curves of the conventional residual test method, where SNR = 20 dB and ANR = 10 dB for both non-stealthy and stealthy attacks are considered. For non-stealthy data attacks, the attacker controls two sensors to inject bad data. From the figure, it can be seen that the conventional method can detect non-stealthy attacks at a successful ratio of around 85 percent given a 10 percent probability of false alarm. However, for stealthy attack, a completely random guess line (the same as coin tossing, i.e., the diagonal line from the bottom left to the top right corner) is obtained, which means the conventional residual test method cannot detect stealthy attacks but always just makes a random guessing decision.

For comparison, based on the same IEEE 9-bus test system, detection performance of the proposed scheme to adaptively detect both non-stealthy and stealthy attacks is demonstrated in Fig. 2b, where the same value of SNR = 20 dB as used in Fig. 2a and various values of ANR are considered. When ANR = 10 dB, compared to the conventional method, the proposed scheme can achieve the same detection probability for detecting non-stealthy attacks, and can significantly improve the detection probability for detecting stealthy attacks. For example, given a 10 percent probability of false alarm, the proposed scheme can successfully detect non-stealthy attacks at a radio of around 86 percent. Different attack levels (i.e., ANR equals 12 dB and 6 dB) are also considered in Fig. 2b. With a higher attack level, whether non-stealthy or stealthy, better detection performance can be achieved using the proposed scheme. It can be anticipated that the proposed scheme is valuable in practical power systems to detect higher-level attacks, since a higher level of attack can always cause larger system state change and is thus more dangerous to the power system operations.

In an IEEE 14-bus test system, there are 20 transmission lines, and thus 34 measurement elements. Figure 3 shows detection performance of the proposed scheme for an IEEE 14-bus test system, where SNR = 20 dB and various values of ANR are considered. For non-stealthy data attacks, we still assume that two sensors are attacked to inject bad data. The proposed scheme can classify and self-adaptively detect both the non-stealthy and stealthy data injection attacks. A random guess line is also shown to present the ROC curve achieved using the conventional method when detecting stealthy attacks. The ROC curves obtained using the proposed scheme are all above the diagonal line, which means that the scheme can achieve very good results for classifying and detecting attacks (significantly better than random guessing). For the stealthy data injection attacks (which are hard to detect using conventional methods), three different levels of ANR are considered. As the attack power level increases, the detection performance of the proposed scheme improves significantly. The proposed scheme also shows very good performance in detecting non-stealthy attacks.

An IEEE 57-bus test system has 80 transmission lines and thus 137 measurement elements. Using the same setting of SNR = 20 dB as used for the IEEE 9-bus and 14-bus systems, detection performance of the proposed scheme for the IEEE 57-bus system is presented in Fig. 4 against various values of ANR. The performance of the conventional method of detecting stealthy data attacks is also shown for comparison. Two specific probabilities of false alarm (30 and 40 percent) are considered. For a given probability of false alarm, at a fixed ANR level, the discrimination threshold τ_1 used for testing non-stealthy data attacks and τ_2 used for testing stealthy data attacks can be computed. Using these thresholds, τ_1 and τ_2, the proposed scheme can classify and detect both the non-stealthy and stealthy attacks. When ANR = 12 dB, to achieve a probability of detection higher than 85 percent for detecting stealthy attacks, we must tolerate a false alarm probability of up to 40 percent. A larger value of ANR leads to a much higher detection probability. As demonstrated in the IEEE 9-bus, 14-bus, and 57-bus test systems, it can be anticipated that the proposed scheme is able to achieve the important objectives of smart grid security in terms of data attack identification and accurate detection.

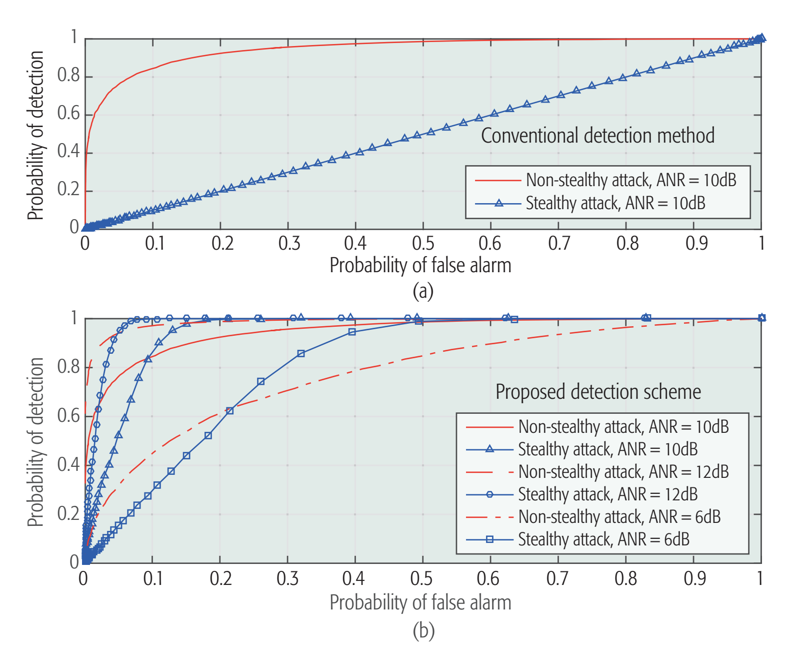


Figure 2. Performance comparison of the proposed detection scheme and conventional detection method for IEEE 9-bus test system, where SNR = 20 dB is considered.

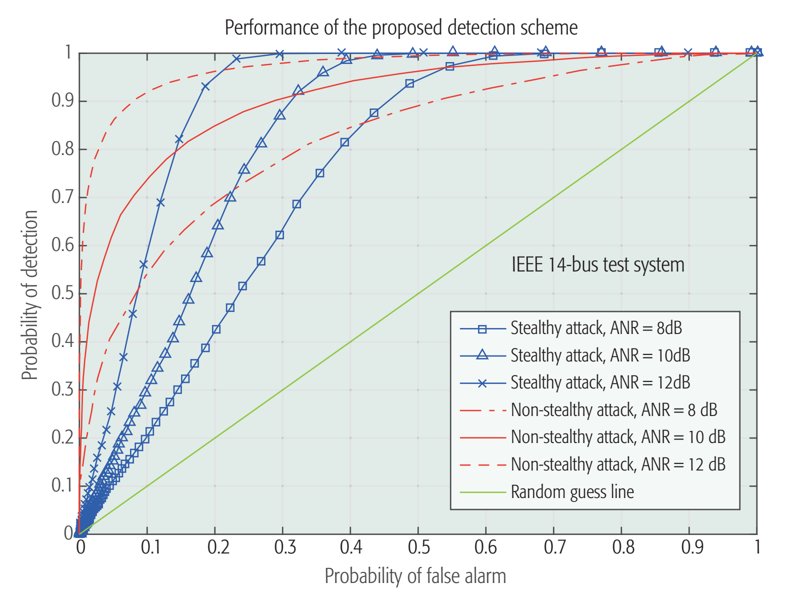


Figure 3. Performance of the proposed detection scheme for IEEE 14-bus test system (SNR = 20 dB and various values of ANR are considered).

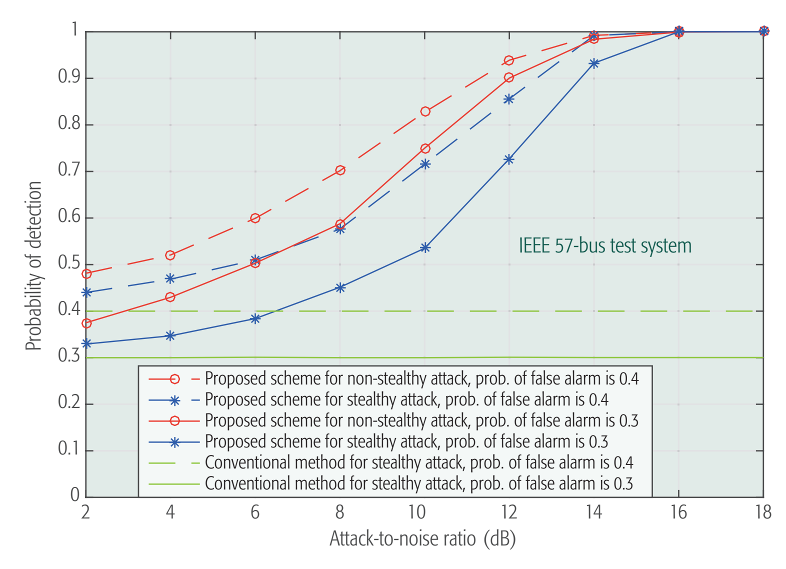


Figure 4. Performance comparison of the proposed detection scheme and conventional detection method for an IEEE 57-bus test system (SNR = 20 dB and various values of ANR are considered).

Conclusions

In this article, we have discussed bad data injection attacks and defense mechanisms in smart grid networks. The problem formulation of state estimation and two types of data injection attacks have been studied. Then, focusing on defense mechanisms, it has been demonstrated that stealthy data attacks are impossible to detect using conventional methods. We have presented a detection scheme that can self-adaptively detect both non-stealthy and stealthy attacks. The scheme comprises determining two estimates of the state of the monitored system using the state measurement data provided by the remote sensing system at two sequential data collection slots, and determining bad data injection attacks by monitoring the measurement variations and state changes between the two slots. Analytical and simulation results have shown that the proposed scheme is efficient in terms of data attack classification and detection accuracy. Once the attack type is known, the power system operators can prioritize their actions or resources to better protect their systems: If the attacks are non-stealthy, the corresponding injected measurements can be removed for another round of state estimation. If the attacks are stealthy, besides removing the measurements, system operators need to change the power network topology, since the measurement matrix has been known to the attacker. Effective action strategies to reduce the risk of being continuously attacked will be left for future work.

Acknowledgment

This work was supported by the European Commission's Horizon 2020 Framework Programme (H2020/2014-2020) under Grant Agreement 646470, SmarterEMC2 Project. This work was also supported by the National Science Foundation under grant CNS-1423408.

References

[1] H. Sun *et al.*, *Smarter Energy: From Smart Metering to the Smart Grid*, ser. Energy Engineering, IET, 2016.

[2] J. Jiang and Y. Qian, "Distributed Communication Architecture for Smart Grid Applications," *IEEE Commun. Mag.*, vol. 54, no. 12, Dec. 2016, pp. 60–67.

[3] H. Sun *et al.*, "Relaying Technologies for Smart Grid Communications," *IEEE Wireless Commun.*, vol. 19, no. 6, Dec. 2012, pp. 52–59.

[4] Y. Yan *et al.*, "A Survey on Smart Grid Communication Infrastructures: Motivations, Requirements and Challenges," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 1, 1st qtr. 2013, pp. 5–20.

[5] S. E. Collier, "The Emerging Enernet: Convergence of the Smart Grid with the Internet of Things," *IEEE Industry Applications Mag.*, vol. 23, no. 2, Mar. 2017, pp. 12–16.

[6] D. B. Rawat and C. Bajracharya, "Detection of False Data Injection Attacks in Smart Grid Communication Systems," *IEEE Signal Processing Letters*, vol. 22, no. 10, Oct. 2015, pp. 1652–56.

[7] Y. Yan *et al.*, "A Survey on Cyber Security for Smart Grid Communications," *IEEE Commun. Surveys & Tutorials*, vol. 14, no. 4, 4th qtr. 2012, pp. 998–1010.

[8] Y. Huang *et al.*, "Real-Time Detection of False Data Injection in Smart Grid Networks: An Adaptive CUSUM Method and Analysis," *IEEE Systems J.*, vol. 10, no. 2, June 2016, pp. 532–43.

[9] J. Hao *et al.*, "Sparse Malicious False Data Injection Attacks and Defense Mechanisms in Smart Grids," *IEEE Trans. Industrial Informatics*, vol. 11, no. 5, Oct. 2015, pp. 1–12.

[10] L. Liu, M. Esmalifalak, and Z. Han, "Detection of False Data Injection in Power Grid Exploiting Low Rank and Sparsity," *2013 IEEE ICC*, June 2013, pp. 4461–65.

[11] S. Cui *et al.*, "Coordinated Data-Injection Attack and Detection in the Smart Grid: A Detailed Look at Enriching Detection Solutions," *IEEE Signal Proc. Mag.*, vol. 29, no. 5, Sept. 2012, pp. 106–15.

[12] M. Esmalifalak *et al.*, "Detecting Stealthy False Data Injection Using Machine Learning in Smart Grid," *IEEE Systems J.*, vol. PP, no. 99, 2014, pp. 1–9.

[13] S. Bi and Y. J. Zhang, "Graphical Methods for Defense Against False-Data Injection Attacks on Power System State Estimation," *IEEE Trans. Smart Grid*, vol. 5, no. 3, May 2014, pp. 1216–27.

[14] T. T. Kim and H. V. Poor, "Strategic Protection Against Data Injection Attacks on Power Grids," *IEEE Trans. Smart Grid*, vol. 2, no. 2, June 2011, pp. 326–33.

[15] N. M. Manousakis and G. N. Korres, "Optimal PMU Placement for Numerical Observability Considering Fixed Channel Capacity — A Semidefinite Programming Approach," *IEEE Trans. Power Systems*, vol. 31, no. 4, July 2016, pp. 3328–29.

Biographies

Jing Jiang (jing.jiang@durham.ac.uk) is a research associate in the School of Engineering and Computing Sciences, Durham University, United Kingdom. In 2011, she obtained her Ph.D. degree from the University of Edinburgh, United Kingdom. During 2011–2014, she was a research fellow with the Centre for Communication Systems Research, University of Surrey, United Kingdom. Her recent research interests include smart grid, next generation wireless communications, massive-MIMO and MIMO techniques, cognitive radio, relay and cooperation techniques, and energy-efficient system design.

Yi Qian (yqian2@unl.edu) is a professor in the Department of Electrical and Computer Engineering, University of Nebraska-Lincoln (UNL). Prior to joining UNL, he worked in the telecommunications industry, academia, and the government. His research interests include information assurance and network security, network design, network modeling, simulations and performance analysis for next generation wireless networks, wireless ad hoc and sensor networks, vehicular networks, smart grid communication networks, broadband satellite networks, optical networks, high-speed networks, and the Internet.

Resilient and Secure Low-Rate Connectivity for Smart Energy Applications through Power Talk in DC Microgrids

Čedomir Stefanović, Marko Angjelichinoski, Pietro Danzi, and Petar Popovski

The authors review power talk, a communication technique specifically developed for direct current MGs, which exploits the communication potential residing within the MG power equipment. Depending on the smart energy application, power talk can be used as either a primary communication enabler or an auxiliary communication system that provides resilient and secure operation.

Abstract

The future smart grid is envisioned as a network of interconnected microgrids (MGs) — small-scale local power networks composed of generators, storage capacities, and loads. MGs bring unprecedented modularity, efficiency, sustainability, and resilience to the power grid as a whole. Due to a high share of renewable generation, MGs require innovative concepts for control and optimization, giving rise to a novel class of smart energy applications, in which communications represent an integral part. In this article, we review power talk, a communication technique specifically developed for direct current MGs, which exploits the communication potential residing within the MG power equipment. Depending on the smart energy application, power talk can be used as either a primary communication enabler or an auxiliary communication system that provides resilient and secure operation. The key advantage of power talk is that it derives its availability, reliability, and security from the very MG elements, outmatching standard off-the shelf communication solutions.

Introduction

The architecture of the power grid has been experiencing a paradigm shift from the classic organization in bulk-generation, transmission, and distribution subsystems into a flexible structure with a high penetration of microgrids (MGs) (Fig. 1). MG is a localized collection of distributed energy resources (DERs), storages, and loads, operating connected to the main grid or in a standalone, islanded mode [1]. The goal is to achieve self-sustainable, efficient, and resilient operation, thereby improving the operation of the entire power network.

Recent advances show that MGs are becoming economically viable [2]. In particular, direct current (DC) MGs are gaining popularity due to the fact that most renewable DERs, storages, and modern loads are DC in nature, implying simpler implementation, reduced costs, higher efficiency, and increased resilience to the main grid disturbances with respect to AC MGs. However, there are several challenges yet to be solved to foster large-scale implementation of DC MGs. From the research and development angle, the major challenge is to incorporate control and communication features pertinent to DC MGs [2,3]. Specifically, renewable DERs show high unpredictability and variability in comparison to traditional bulk generation, requiring novel control and optimization approaches, at both the intra- and inter-MG levels.

DC MG control architecture is organized in primary, secondary, and tertiary layers [3]. The primary control is the fastest, operating in the frequency range of 0.1–1 MHz, regulating the MG voltage and power flow such that high frequency load/generation variations are compensated. It is implemented in a decentralized manner, executed according to the measurements locally available to DERs and the control references provided by the tertiary control. The secondary control operates with frequencies in the range of 1–10 kHz, and its task is to eliminate the steady state voltage drift and power sharing mismatch, introduced by the primary control due to its decentralized nature. Finally, the tertiary control comprises smart energy applications that minimize power dissipation losses and generation costs, as well as maximize the economic viability of the system by providing the optimal references for the primary control. It is the slowest control level, running periodically every 5–30 minutes. Typical smart energy applications include optimal power flow (OPF), optimal economic dispatch (OED), demand-response (DR), unit commitment (UC), and so on.

Unlike the primary control layer, the secondary and tertiary control layers require information that is not available via local measurements. Depending on the control application, this information may comprise voltage/current measurements at remote DERs, instant DER generation capacities, loads' demands and admittance matrix, control directives, information on generation costs, ramp-up constraints, and so on [3]. In other words, secondary and tertiary MG control applications require communication support. The actual amount of data that should be communicated is small, as it is typical for machine-to-machine communications, and the periodicity of the communication exchanges should follow the periodicity (i.e., frequency) of the control application.

Information exchange is also required to enable higher level inter-MG operation, executed

The authors are with Aalborg University.

Digital Object Identifier: 10.1109/MCOM.2017.1700178

0163-6804/17/$25.00 © 2017 IEEE

Secondary and tertiary MG control applications require communication support. The actual amount of data that should be communicated is small, as it is typical for machine-to-machine communications, and the periodicity of the communication exchanges should follow the periodicity (i.e., frequency) of the control application.

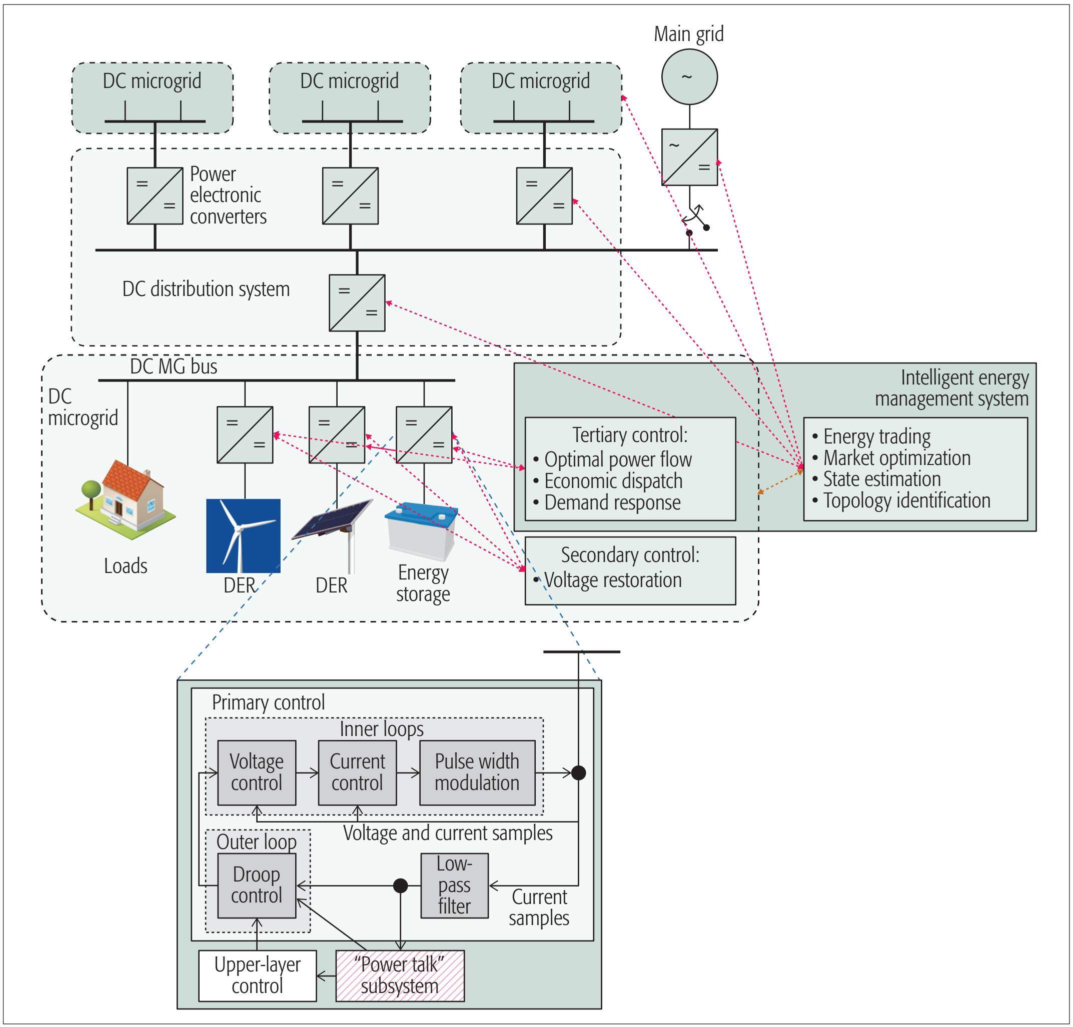


Figure 1. Cluster of DC microgrids: DC MGs are characterized by extensive use of power electronic converters that regulate the power generation, executing intra- and inter-MG control algorithms.

within and among MG clusters (Fig. 1). In fact, modern MGs have the capability of integrating with the existing power system and dealing with bidirectional exchange of power, thus increasing the overall grid availability in case of fault events and reducing the stress on overloaded portions of the system [4]. Moreover, the interaction of higher-level control applications enables state estimation, topology identification, energy trading, and market optimization. These higher-level MG control applications, including tertiary MG control applications, form an intelligent energy management system (IEMS) (Fig. 1), which supervises and optimizes the overall MG operation in a broader environment where MGs are placed [5], like commercial buildings and residential blocks.

To meet the communication needs of intra- and inter-MG control and optimization, a standard approach is to employ an external communication solution, such as wireless or power line communications (PLC) [3, 4, 6]. However, relying fully on external communication systems may compromise the goal of self-sustainable and resilient MG operation, due to their limited availability, reliability, and security [7]. In this article, we propose a resilient and secure communication framework for DC MGs, called power talk, which exploits only the communication potential residing within the MG power equipment. Depending on the smart energy application, power talk can be used as either a primary communication enabler or an auxiliary communication system that fosters resilient and secure operation. Besides the fact that power talk derives its availability, reliability, and security from the MG components, it also does not require installation of any additional hardware and provides complete coverage of the MG system.

Basics of Power Talk

MGs are characterized by extensive use of power electronic converters (PECs), which are digital signal processors interfacing DERs and storages to the buses [2, 3] (Fig. 1). In a DC MG, a PEC (i.e., the DER it controls) can operate as either a voltage source converter (VSC) or a current source converter (CSC). When operating in VSC mode, a PEC participates in MG control and optimization through regulation of the output power of its DER. In CSC mode, a PEC does not participate in the MG control and optimization, and the DER it controls generates the maximum output power. Also, a PEC can change its operating mode from VSC to CSC and vice versa as needed.

The block diagram of a PEC operating as a VSC can be seen in Fig. 1. A VSC constantly

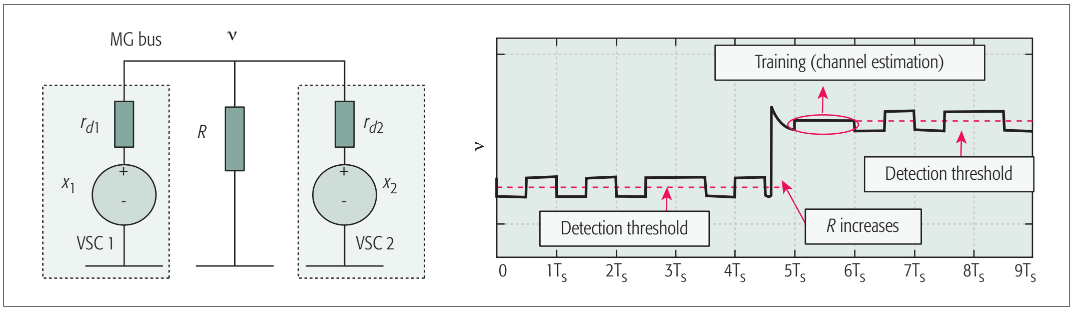


Figure 2. a) example of a single-bus MG; b) communication from VSC 1 to VSC 2 via bus voltage deviations and the effects of the load change.

Power talk is designed with the aim of establishing a general digital interface among PECs in MG that can be used for any control and optimization application. It is envisioned as an upgrade of the control functionality of PECs with communication capabilities, without using any additional communication hardware.

measures bus voltage and current with switching frequency that ranges from several tens of kilohertz to a couple of megahertz, and, based on the measurements, executes the primary control algorithm that is commonly implemented in the form of droop control [3]. VSCs also perform the upper layer control and optimization functions, which require communication support.

Being digital signal processors, PECs can also engage in communication-related tasks using the buses interconnecting them as the communication medium. We illustrate this through a simple example of a single-bus MG, depicted in Fig. 2, with two PECs that operate as droop controlled VSCs and a single resistive load *R*. The steady-state bus voltage *v*, observed by both VSCs, is given by

$$v = \frac{R(r_{d1}x_2 + r_{d2}x_1)}{R(r_{d1} + r_{d2}) + r_{d1}r_{d2}} \tag{1}$$

where x_1/x_2 is the reference voltage and r_{d1}/r_{d2} is the virtual resistance of VSC1/VSC2, which are controllable droop parameters. Obviously, if VSC1 deviates its x_1 and/or r_{d1}, this will cause deviations of *v*. From the communication engineering point of view, x_1 and r_{d1} can be seen as the *inputs*, and *v* as the *output* of the communication channel between VSC1 and VSC2. This simple but fundamental insight can be exploited to design a communication system among PECs in the MG, which is embedded in the primary control and uses steady-state deviations of the bus voltages to transfer information. This is the key concept of power talk, which can be implemented via a software modification of PEC architecture.

The idea of using MG bus for communication among PECs was proposed in several works that target predefined MG setups and control applications, in which PECs perform control actions based on the observation of the bus voltage [8, references therein]. Establishment of a low-rate communication interface over DC bus by selecting predefined PEC switching frequencies was proposed in [9], and using pulse-width modulation in [10]. Both [9, 10] address only physical layer aspects, neglecting the functionalities needed for setting up fully operational communication links. In contrast, power talk is designed with the aim of establishing a general digital interface among PECs in MG that can be used for any control and optimization application. Finally, we remark that power talk uses power lines to convey information like in PLC. However, all PLC standards require installation of communication modems, whereas power talk is envisioned as an upgrade of the control functionality of PECs with communication capabilities, without using any additional communication hardware.

Main Characteristics of Power Talk

To create a functional communication solution through primary control, there are several important aspects to be taken into account. We elaborate them in a general multibus MG setup (Fig. 3).

Signaling rate: An MG bus typically requires 1–10 ms to reach a steady state, implying that signaling rates in power talk are on the order of 100 Bd–1 kBd. These rates are adequate for all IEMS applications, but not for secondary control.

Synchronization: Virtually all MG control applications are periodic in nature, such that power talk should be invoked at regular intervals. Further, being a baseband digital communication technique, power talk also requires packet-level and symbol-level synchronization. If PECs are equipped with an external synchronization interface, like GPS, the synchronization may easily be achieved. Otherwise, PECs can rely on their internal clocks for coarse synchronization, and then apply standard techniques to achieve and maintain precise packet- and symbol-level synchronization, such as use of synchronization preambles and adequate signaling formats [11].

Multiple access: Power talk establishes a multiple access communication channel. This can be observed in the context of the simple example depicted in Fig. 2, for which the steady-state value of the bus voltage, given by Eq. 1, simultaneously depends on the control parameters of both VSC 1 and VSC 2, which are x_1, r_{d1} and x_2, r_{d2}, respectively. In other words, the output of the communication channel simultaneously depends on all its inputs. The same can be shown to hold in the general model, depicted in Fig. 3. Considering the expected periodicity of power talk sessions and static/slowly changing MG control configuration (i.e., assignment of PEC operating modes), a straightforward approach is to employ time-division multiplex (TDM) [12], thus imposing half-duplex channels. Another appealing option, motivated by the fact that the set of transmitters is known a priori, is to use coding strategies for multiple access. In this approach, all VSCs simultaneously exchange information among themselves [12], achieving all-to-all full duplex communication.

Channel state information: Application of Kirchhoff's laws reveals that the steady-state bus voltage v_n that VSC *n* observes (Fig. 3) depends

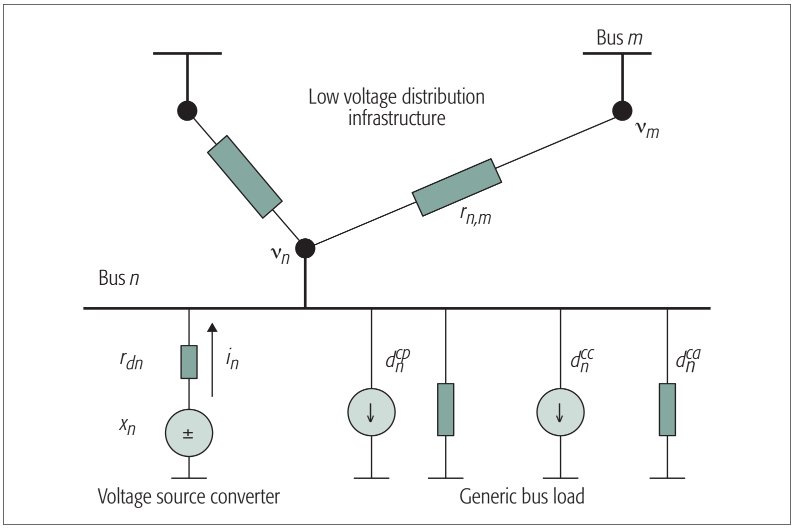


Figure 3. General electrical model of a multibus MG. Per bus, there is a single VSC and an aggregate load representation with a constant power, a constant current, and a resistive component, with power demands d^{cp}, d^{cc}, and d^{ca} at rated bus-voltage x_R, respectively; the constant power component of the load also accounts for the potential CSCs in the bus. r_{nm} denotes the resistance of the line connecting buses n and m.

on values of all reference voltages x, virtual resistances r_d, line resistances r, and load components d^{cp}, d^{cc}, and d^{ca} in the system [12]. In other words, the state of the communication channel is determined by the values of all components of the electrical model of the MG. The knowledge of these values is unavailable a priori, necessitating a training phase in which receiving VSCs learn the states of the channels they observe before engaging in communications. The training can be done through coordinated actions of all VSCs in the system [12]. This is reminiscent of standard approaches in wireless communication systems where the channels are subject to random behavior, and some form of training is required for channel estimation.

Load changes: Loads in MGs change sporadically, but unpredictably. If a load change occurs during a power talk session, the channel state information becomes invalidated, which may require restart of training phase and of information transfer, as illustrated in Fig. 2b. Load changes may be detected using standard error detection methods on the physical layer (e.g., using CRC codes).

Noise: The observations of bus voltages contain measurement noise, which can be modeled as additive, Gaussian, and white, where the typical values for the standard deviation of the voltage measurement noise (in volts per sample) in low voltage distribution systems are in the range of 0.01–0.1 percent of the voltage rating per unit [12]. The signal-to-noise ratio (SNR) of power talk is determined by the amplitude of the allowed bus-voltage deviations used for power talk that the MG can tolerate, and the noise power after averaging of the samples during a signaling interval. In this respect, it can be shown that power talk operates in a very high SNR regime [12], such that the impact of noise is rather small. Finally, if required, the impact of noise can be further reduced using standard channel coding methods.

Electrical constraints: Virtual resistances and reference voltages by default feature constraints on their minimal and maximal values. Moreover, information-carrying deviations of these control parameters incur power deviations on the buses, and should be as small as possible with respect to the optimal power levels prescribed by smart energy applications. In terms of communication system design, these constraints define the signaling space of power talk, in which one can construct optimized symbol constellations [13].

Security: Power talk, like PLC in general, offers security advantages with respect to the use of wireless networking for the MG information exchange, as the access to the communication channel can be made only by the devices that are physically attached to the MG power lines. Thus, power talk is inherently more resilient and secure than wireless solutions, to both passive attacks (eavesdropping, jamming) and active attacks (man the middle, denial of service, etc.). With respect to PLC, power talk has the advantage that the communication is directly actuated by the PEC control software without delegating it to an external modem. In contrast, PLC requires establishment of a trustful relationship of the MG control layer with the external communication network.

The Architecture of Smart Energy Applications with Power Talk

The proposed functional architecture of PECs executing smart energy applications with power talk is depicted in Fig. 3. The power talk block provides information exchanges for the IEMS, that is, for all tertiary and inter-MG control and optimization applications, as these have slow dynamics and require modest data rates. On the other hand, the frequency with which the secondary control operates is beyond reach of power talk, mandating use of an external, high-rate communication interface. Nevertheless, the power talk interface can be used as an auxiliary channel for the exchange of the information related to the state of the external network. The examples are exchanges of connectivity status and alarm messages that can be used for the external network reconfiguration [14] or for establishment of a security context that can be used by the external network [15].

In the following, we illustrate the potential of power talk via two example case studies. In the first, power talk is used as a communication solution for the tertiary control application of the optimal economic dispatch, while in the second, power talk is used in the context of the distributed secondary control as an auxiliary channel for the reconfiguration of the external wireless network under jamming attack.

Case Study 1: Optimal Economic Dispatch

In MGs that are predominantly based on renewable technologies, the IEMS collects information about the generation capacities and runs OED periodically (e.g., every 5–30 minutes). The goal of OED is to dispatch the VSCs based on the instant generation capacities, such that the total generation cost is minimized and the load is balanced.

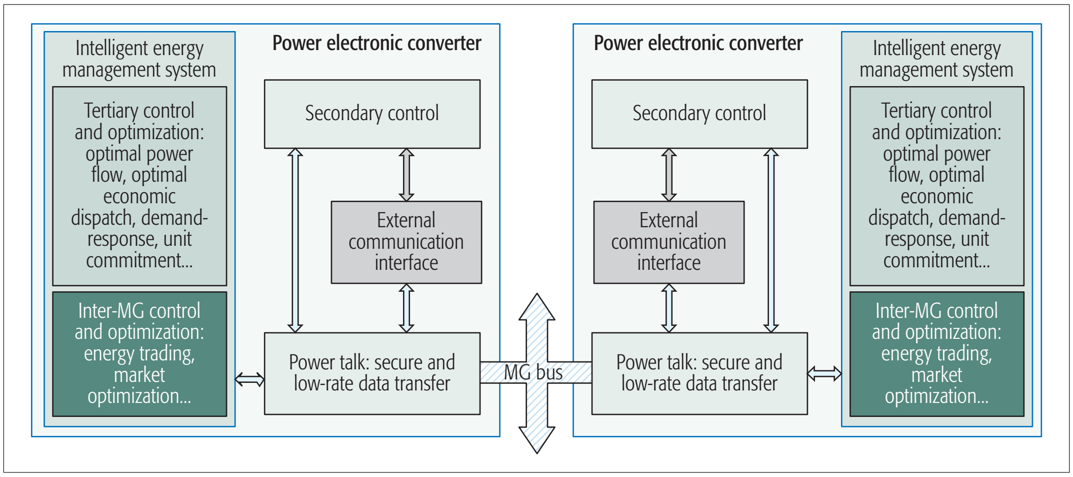

Intelligent energy management system
Tertiary control and optimization: optimal power flow, optimal economic dispatch, demand-response, unit commitment...
Inter-MG control and optimization: energy trading, market optimization...
Power electronic converter
Secondary control
External communication interface
Power talk: secure and low-rate data transfer
MG bus
Power electronic converter
Secondary control
External communication interface
Power talk: secure and low-rate data transfer
Intelligent energy management system
Tertiary control and optimization: optimal power flow, optimal economic dispatch, demand-response, unit commitment...
Inter-MG control and optimization: energy trading, market optimization...

Figure 4. The proposed functional architecture of PEC executing smart energy applications with power talk.

We design a simple power talk protocol to support distributed optimal economic dispatch. The protocol consists of periodic power talk phases, each phase preceding the next optimal dispatch period, during which DERs exchange the information required by the optimal dispatch.

We focus on distributed OED (DOED) implementation with linear cost functions, typically used for renewable generation. The MG hosts U dispatchable VSCs, with generation capacities denoted by $p_{u,\max}$. Each VSC is assigned incremental cost c_u per unit of generated power; without loss of generality the costs follow the ordering $c_1 \leq c_2 \leq \ldots \leq c_U$. The load demand is denoted by d. A typical situation in which d is known a priori is assumed (e.g., via an accurate forecasting performed one day in advance). In the distributed implementation, VSC u, besides d, needs to know $p_{k,\max}$ for each k that satisfies $c_k \leq c_u$.

We design a simple power talk protocol to support DOED. The protocol consists of periodic power talk phases, each phase preceding the next DOED period, during which DERs exchange the information required by DOED. A power talk phase consists of N time slots, each with duration of T_S seconds. The power talk phase uses time-division multiple access (TDMA): the slots are divided into U consecutive sub-phases, and each sub-phase is assigned to one of the VSCs and consists of Q slots, such that N = QU, as depicted in Fig. 5a. VSC u quantizes its generation capacity $p_{u,\max}$ into a binary string of Q bits, which is then transmitted in the dedicated sub-phase via uncoded binary amplitude modulation of the reference voltage x_u. Specifically, a logical 0/1 is transmitted by deviating the reference voltage x_u from its nominal value $x_{u,\text{nom}}$ (which is determined by DOED) by a predefined deviation amplitude $-\lambda/\lambda$, respectively (i.e., $x_u = x_{u,\text{nom}} \pm \lambda$). The receiving VSCs simply compare the bus-voltage level observed in each slot to the bus-voltage level prior to the power talk phase, thereby detecting the transmitted bits.

At the end of the power talk phase, each VSC acquires the knowledge of the instant generation capacities. However, this knowledge is imperfect due to quantization and noise induced detection errors. Thus, the resulting dispatch policy in the next DOED period might be suboptimal, leading to an increase of the generation cost in comparison to the optimal policy. Another type of cost that should be taken into account is the cost of the power deviations due to information-carrying bus-voltage deviations in the power talk phase. These two costs form the total cost incurred by power talk.

We instantiate the proposed protocol in a single-bus MG with U = 6 VSCs, where the rated voltage of the bus is x_R = 48 V, the sampling (switching) frequency is 50 kHz, the standard deviation of the converters sampling noise is 0.05 V/sample, and the duration of power talk slots is T_S = 5 ms. The samples obtained in each slot are averaged, which implies that the standard deviation of the noise after the averaging is reduced to roughly 0.0032 V in a slot. Finally, the value of a bit transmitted in a slot is decided by comparing the produced average value of the bus voltage samples in the slot with the nominal bus voltage level that is determined by DOED.

We measure the efficiency of the proposed approach via the average relative increase of the generation cost when power talk is used with respect to the generation cost when an ideal, "costless" communication solution is used, denoted by δ. Figure 5b depicts δ as function of the number of quantization bits Q, parameterized with reference voltage deviation amplitudes δ in the range 0.02–1 V (corresponding to SNR in the range of 16–50 dB and average power deviation in the range of 5–200 W). We observe that the largest values of δ occur for very small values of Q, as in this case the received information about the generation capacities is very imprecise. On the other hand, for $Q > 5$, the generation cost increase becomes dominated by the power spent on the power talk phase. In this example, δ is minimized for $Q = 4$, proving that the length of the messages in smart energy applications is indeed very short. Finally, we note that the minimal relative cost increase is below 1 percent, making power talk a viable candidate in comparison to solutions that employ external communication systems, which involve costs of installation, maintenance, and operation.

Case Study 2: Robust and Secure Distributed Secondary Control

An envisioned application for low-voltage MGs is the enforcement of the power reliability of critical buildings (e.g., commercial buildings), in which unexpected voltage fluctuations may cause damage to the electronic equipment. In this scenario, the MG is composed of a high number of small DERs that are networked by short-range wireless

We have presented two case studies that clearly show the utility and the potential of power talk, one related to economic dispatch and the second to the cybersecurity in MGs. Future work includes integration of power talk in other smart energy applications and processes and co-design of power talk with the distributed control algorithms.

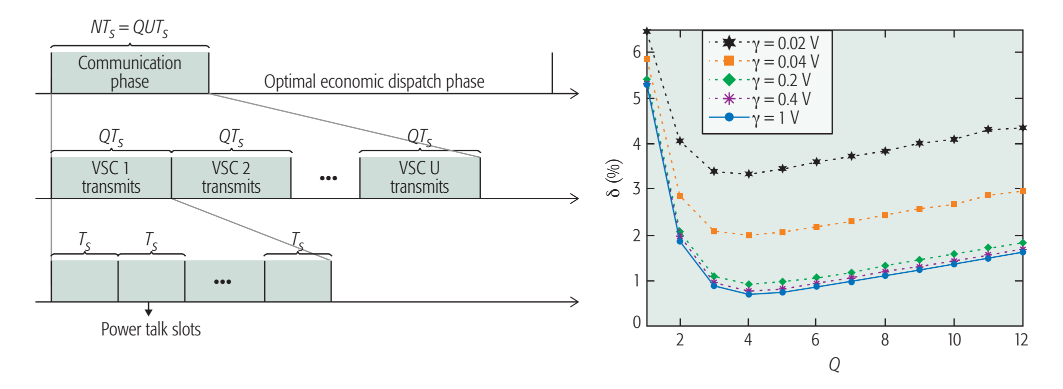


Figure 5. a) Temporal organization of the proposed protocol; b) the relative cost increase when power talk is used in the communication phase of DOED, compared to the ideal, costless communication solution.

communication interfaces (e.g., IEEE 802.11), as depicted in Fig. 6a. The voltage restoration is supported by distributed algorithms that, contrary to the centralized control approach, permit an easy network reconfiguration, enhancing the grid scalability and relieving it from a single point of failure [3]. The secondary control is executed by a subset of active DERs (i.e., VSC units), while the others work as CSCs in order to maximize the overall generation.

The communication graph of the networked VSCs should be strongly connected to enable proper execution of the secondary control. However, adverse channel conditions, such as continuous jamming, may cause the graph disconnection and the formation of insulated subsets of VSCs. In this case, a consensus-based secondary control is prevented from converging to a global solution, reflecting the physical effect of unbalanced power sharing among DERs [7].

A possible approach to deal with this scenario is to select a new subset of voltage regulators, by switching some of the CSCs to VSC operation mode such that the communication graph becomes connected again, while the insulated VSCs are switched to CSC mode. The proposed network reorganization can be done via periodically invoked power talk sessions, similar to the approach outlined in the previous case study. Specifically, the proposed protocol adopts the following steps:

- All DERs broadcast wireless packets.
- Based on the received wireless broadcasts, each DER broadcasts a list of reachable neighbors and its current power generation capacity over the power talk channel in a TDMA fashion, where it is assumed that CSCs temporarily switch to VSC mode in order to participate in power talk communication.
- Finally, each DER locally decides on its operation mode — VSC or CSC, such that the communications graph of the wireless network is connected, the voltage restored, and the power sharing balanced.

The last step is possible due to the fact that all DERs now share the same knowledge. In an unfavorable case in which jamming is such that it is not possible to wirelessly network VSCs to facilitate adequate voltage restoration, the use of power talk enables dissemination of the information that can be used to detect such an event.

We instantiate the proposed approach via the example MG depicted in Fig. 6a. The electrical part of the scheme is simulated in Simulink/PLECS, with the rated of voltage 48 V and with the variant of the power talk introduced in case study 1 with T_S = 2.5 ms, Q = 8, and λ = 0.25 V. The wireless communication part of the scheme adopts IEEE 802.11-n. The MG is composed by 9 DERs in which DERs 2, 5, 6, and 9 are initially participating in the secondary control. Their connectivity is undermined by a jamming device placed in proximity of DER 5, which is capable of continuous transmission that prevents the communication. When a load variation occurs, such as the activation of a resistive load at t = 7 s in Fig. 6b, the absence of global connectivity reflects in a current imbalance. This is detected in the consecutive power talk phase, and a reorganization is triggered. The result is the formation of a new secondary control set, composed by DERs 1, 2, 7, and 10. In conclusion, placing power talk side by side with the high-bandwidth, but at the same time unreliable, wireless network, and exploiting the cyber-physical properties of the MG is a viable and promising solution to increase the system robustness.

Conclusions

This article has reviewed the use of power talk in DC microgrids, a low-rate communication technique that reuses the power electronics and does not rely on dedicated communication hardware. The use of power talk has both architectural and functional implications on the operation of a system of MGs and can support multiple smart energy applications. Despite the low rate, the reliability and security of the low-rate communication channel offered by power talk can have a significant impact on the overall performance. We have presented two case studies that clearly show the utility and the potential of power talk, one related to economic dispatch and the second to the cybersecurity in MGs. Future work includes integration of power talk in other smart energy applications and processes, and co-design of power talk with the distributed control algorithms.

Acknowledgment

The work presented in this article was supported in part by EU, under grant agreement no. 607774 "ADVANTAGE."

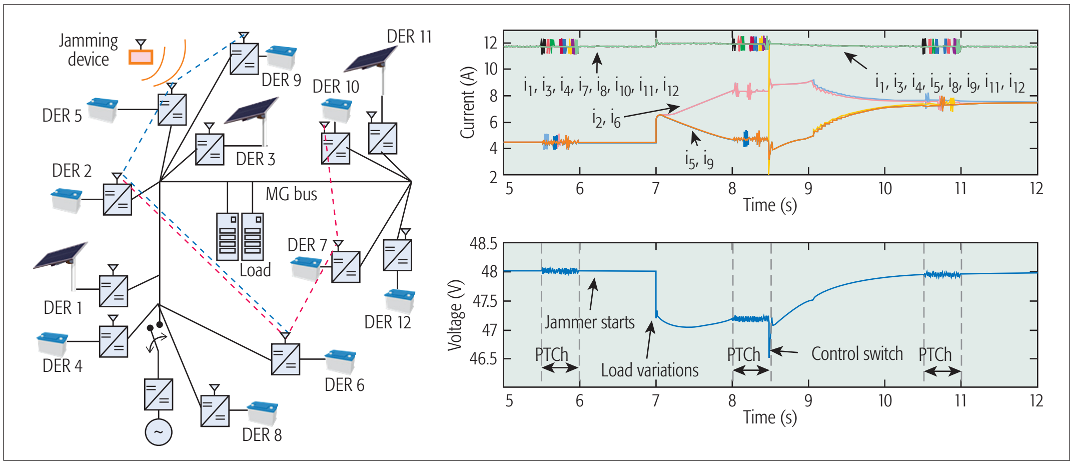


Figure 6. Left: the MG considered in the study case. The dashed blue lines represent the links before the reconfiguration, the dashed red the communication graph after the activation of the jammer (located in the upper-left corner); right: PLECS simulation of the secondary control reconfiguration in case of jamming attack. We report the output current of each DER and the voltage measured on the MG bus. Observe the periodic power talk channels (PTCh) used to signal the network information.

References

[1] R. H. Lasseter, "MicroGrids," *Proc. IEEE Power Eng. Soc. Winter Mtg.*, New York, NY, Jan. 2002.

[2] L. E. Zubieta, "Are Microgrids the Future of Energy?: DC Microgrids from Concept to Demonstration to Deployment," *IEEE Electrific. Mag.*, vol. 4, no. 2, May 2016, pp. 37–44.

[3] T. Dragicevic *et al.*, "DC Microgrids; Part I: A Review of Control Strategies and Stabilization Techniques," *IEEE Trans. Power Electron.*, vol. 31, no. 7, July 2016, pp. 4876–91.

[4] S. Moayedi and A. Davoudi, "Distributed Tertiary Control of DC Microgrid Clusters," *IEEE Trans. Power Electron.*, vol. 31, no. 2, Feb. 2016, pp. 1717–33.

[5] P. Palensky and D. Dietrich, "Demand Side Management: Demand Response, Intelligent Energy Systems, and Smart Loads.," *IEEE Trans. Industrial Informatics*, vol. 7, no. 3, Aug. 2011, pp. 381–88.

[6] S. Galli, A. Scaglione, and Z. Wang, "For the Grid and Through the Grid: The Role of Power Line Communications in the Smart Grid," *Proc. IEEE*, vol. 99, no. 6, June 2011, pp. 998–1027.

[7] P. Danzi *et al.*, "On the Impact of Wireless Jamming on the Distributed Secondary Microgrid Control," *Proc. IEEE GLOBECOM 2016, Wksp. Cyber-Physical Smart Grid Security and Resilience*, Washington, DC, Dec. 2016.

[8] T. Dragicevic, J. M. Guerrero, and J. C. Vasquez, "A Distributed Control Strategy for Coordination of an Autonomous LVDC Microgrid Based on Power-Line Signaling," *IEEE Trans. Industrial Electron*, vol. 61, no. 7, July 2014, pp. 3313–26.

[9] Z. Lin *et al.*, "Novel Communication Method Between Power Converters for DC Microgrid Applications," *Proc. IEEE ICDCM 2015*, Atlanta, GA, June 2015.

[10] J. Wu *et al.*, "Power Conversion and Signal Transmission Integration Method Based on Dual Modulation of DC-DC Converters," *IEEE Trans. Industrial Electron.*, vol. 62, no. 2, Feb. 2015, pp. 1291–1300.

[11] M. Angjelichinoski *et al.*, "Power Talk: How to Modulate Data over a DC Micro Grid Bus using Power Electronics," *Proc. IEEE GLOBECOM 2015*, San Diego, CA, Dec. 2015.

[12] M. Angjelichinoski *et al.*, "Multiuser Communication Through Power Talk in DC Microgrids," *IEEE JSAC*, vol. 34, no. 7, July 2016, pp. 2006–21.

[13] M. Angjelichinoski *et al.*, "Power Talk in DC Micro Grids: Constellation Design and Error Probability Performance," *Proc. IEEE SmartGridComm 2015*, Miami, FL, Nov. 2015.

[14] P. Danzi *et al.*, "Anti-Jamming Strategy for Distributed Microgrid Control Based on Power Talk Communication," *Proc. IEEE ICC 2017 Wksp. Integrating Commun., Control, and Computing Technologies for Smart Grid*, Paris, France, May 2017.

[15] M. Angjelichinoski *et al.*, "Secure and Robust Authentication for DC MicroGrids based on Power Talk Communication," *Proc. IEEE ICC 2017*, Paris, France, May 2017.

Biographies

Ĉedomir Stefanović [S'04, M'11, SM'17] received his Dipl.-Ing., Mr.-Ing., and Ph.D. degrees in electrical engineering from the University of Novi Sad, Serbia. He is currently an associate professor at the Department of Electronic Systems, Aalborg University, Denmark. In 2014 he was awarded an individual postdoctoral grant by the Danish Council for Independent Research (Det Frie Forskningsråd). His research interests include communication theory, and wireless and smart grid communications.

Marko Angjelichinoski [S'15] received his B.Sc. and M.Sc. degrees in telecommunications from Ss. Cyril and Methodius University, Skopje, Macedonia, in 2011 and 2014, respectively. Presently he is a Ph.D. fellow at Aalborg University. His research interests are in the areas of statistical signal processing, estimation, detection, and information theory with applications in next generation systems such as cognitive radio networks and smart grid.

Pietro Danzi [S'16] is a doctoral student of wireless communications at Aalborg University. He also has a Marie Curie fellowship as an Early Stage Researcher. Previously, he obtained his M.Sc. degree in telecommunication engineering from Università degli Studi di Padova, Italy, with a thesis on algorithms for pattern selection of reconfigurable antennas. His current interests include machine-type communication protocols, blockchain protocols, and cyber-security for smart grids.

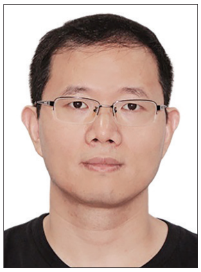
Petar Popovski (S'97, A'98, M'04, SM'10, 785 F'16) is a professor at Aalborg University. He received his Dipl.-Ing./Magister Ing. in communication engineering from Sts. Cyril and Methodius University, and his Ph.D. from Aalborg University. He received an ERC Consolidator Grant (2015) and the Danish Elite Researcher award (2016). He is an Area Editor for *IEEE Transactions on Wireless Communications*. His research interests are in wireless communications/networks and communication theory.

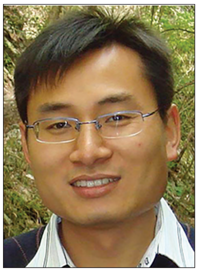
BEHAVIOR RECOGNITION BASED ON WI-FI CSI: PART 1

Bin Guo

Yingying (Jennifer) Chen

Nic Lane


Yunxin Liu


Zhiwen Yu

Human behavior recognition is the core technology that enables a wide variety of human-machine systems and applications (e.g., healthcare, smart homes, and fitness tracking). Traditional approaches mainly use cameras, radars, or wearable sensors. However, all these approaches have certain disadvantages. For example, camera-based approaches have the limitations of requiring line of sight with enough lighting, potentially breaching human privacy. Low-cost radar-based solutions have limited operation range of just tens of centimeters.

Recently, Wi-Fi channel state information (CSI)-based human behavior recognition approaches are attracting increasing attention. The rationale is that different human behaviors introduce different multi-path distortions in Wi-Fi CSI. Compared to traditional approaches, the key advantages of Wi-Fi CSI-based approaches are that they do not require lighting, provide better coverage as they can operate through walls, preserve user privacy, and do not require users to carry any devices as they rely on the Wi-Fi signals reflected by humans. As a result, the recognition of quite a number of behaviors that are difficult based on traditional approaches have now become possible, including fine-grained movements (e.g., gesture and lip language), keystrokes, drawings, gait patterns, vital signals (e.g., breathing rate), and so on. However, Wi-Fi CSI-based behavior recognition still faces a number of challenges: What are the fundamental theories and models that can steer the development of accurate, robust, and fine-grained CSI sensing systems? How do we overcome the impact of noise and ensure the performance of CSI-enabled systems? How do we recognize the behavior of multiple users?

This Feature Topic provides an opportunity for researchers and product developers to review and discuss the state of the art and trends of CSI-based behavior recognition techniques. A total of 16 articles were submitted from around the globe via the open call. In order to ensure high reviewing standards, three to four reviewers evaluated each article. The finally accepted articles are organized into Part 1 and Part 2. In Part 1 (this issue), you can find four of them as follows; the other ones will be published as Part 2 in a later issue. The selected articles cover different topics, such as literature review, pattern/model-based recognition approaches, and novel applications.

The first article, "Device-Free WiFi Human Sensing: From Pattern-Based to Model-Based Approaches," by Wu *et al.*, reviews the research in Wi-Fi CSI-based device-free human behavior sensing in recent years. The authors point out the research trend that would evolve from pattern-based to model-based approaches, and suggest that researchers leverage the Fresnel zone model as the basis of wireless human sensing and extend it to a general sensing model.

In the second article, "A Survey on Behavior Recognition Using WiFi Channel State Information," Yousefi *et al.* present a survey of recent techniques for human behavior recognition using channel information of Wi-Fi devices. They then show the performance of deep learning techniques such as LSTM for supervised classification of user activities. Finally, the authors discuss the challenges in activity recognition using Wi-Fi CSI and suggest research directions for future study.

The third article, "Wi-Fi Radar: Recognizing Human Behavior with Commodity Wi-Fi" by Zou *et al.*, first defines Wi-Fi radar, which is a novel kind of system based on commodity Wi-Fi. By surveying the latest works, the authors summarize the general framework of existing Wi-Fi radar systems, and figure out that the design of these systems mainly follows a data-driven approach or a model-based approach. For each kind of Wi-Fi radar, the article gives a detailed introduction to the fundamental principles and state-of-the-art applications.

The fourth article, "Human Behavior Recognition Using Wi-Fi CSI: Challenges and Opportunities" by Chen *et al.*, provides a tutorial on human behavior recognition (HBR) using Wi-Fi CSI. The article first reviews the state of the art of HBR, based on the techniques that have driven recent progress. It then provides insights on the future directions of HBR research.

In concluding this overview, we would like to address our special thanks to Dr. Osman Gebizlioglu, the Editor-in-Chief of *IEEE Communications Magazine*, and Jennifer Porcello and Peggy Kang for their great support and effort throughout the whole publication process of this Feature Topic. We are also grateful to all the authors for submitting their papers and the reviewers for their professional and timely work in making it possible to publish this Feature Topic.

BIOGRAPHIES

BIN GUO [SM] (guobin.keio@gmail.com) is currently a professor at Northwestern Polytechnical University, China. He received his Ph.D. degree in computer science from Keio University, Tokyo, Japan, in 2009. His research interests include ubiquitous computing and mobile crowdsensing. He has served as an Associate Editor of *IEEE Communications Magazine* and *IEEE Transactions on Human-Machine Systems*.

YINGYING (JENNIFER) CHEN is a tenured professor at Stevens Institute of Technology. She leads the Data Analysis and Information Security (DAISY) Lab and is also the graduate program director of Information and Data Engineering and Networked Information Systems. Her research interests include smart healthcare, the Internet of Things, and mobile sensing.

NIC LANE is a senior lecturer at University College London and a principal scientist at Nokia Bell Labs. He received his Ph.D. degree from Dartmouth College. Before joining Nokia Bell Labs, he spent four years as a lead researcher at Microsoft Research based in Beijing. His research interests include mobile computing and deep learning.

YUNXIN LIU [SM] is a researcher in the System Research Group, Microsoft Research Asia. He received his Ph.D. in computer science from Shanghai Jiao Tong University. His research interests are mobile systems and networking.

ZHIWEN YU [SM] is currently a professor at Northwestern Polytechnical University, China. He worked as an Alexander Von Humboldt Fellow at Mannheim University, Germany, from November 2009 to October 2010. His research interests cover ubiquitous computing and HCI.

Device-Free WiFi Human Sensing: From Pattern-Based to Model-Based Approaches

Dan Wu, Daqing Zhang, Chenren Xu, Hao Wang, and Xiang Li

ABSTRACT

Recently, device-free WiFi CSI-based human behavior recognition has attracted a great amount of interest as it promises to provide a ubiquitous sensing solution by using the pervasive WiFi infrastructure. While most existing solutions are pattern-based, applying machine learning techniques, there is a recent trend of developing accurate models to reveal the underlining radio propagation properties and exploit models for fine-grained human behavior recognition. In this article, we first classify the existing work into two categories: pattern-based and model-based recognition solutions. Then we review and examine the two approaches together with their enabled applications. Finally, we show the favorable properties of model-based approaches by comparing them using human respiration detection as a case study, and argue that our proposed Fresnel zone model could be a generic one with great potential for device-free human sensing using fine-grained WiFi CSI.

INTRODUCTION

Radio-based human behavior sensing has become an active research area due to the pervasiveness of such wireless signals. While most existing work focuses on device-based scenarios,[1] recently device-free sensing solutions have been increasingly popular because they significantly improve the usability and practicality of indoor applications, such as intrusion detection, elder care, and healthcare. The earliest work on device-free human sensing, called "sensorless sensing," was introduced by Woyach *et al.* in 2006 [1]. By observing human motion and the resultant signals in a wireless sensor network, Woyach *et al.* noticed that the motion of a human can cause a series of signal fading spots, and further demonstrated the possibility and promise of using wireless sensors for human presence detection in a contact-free manner. Soon after that, in 2007, Youssef *et al.* [2] experimentally verified that human motion causes variations on the received signal strength indicator (RSSI), and simple features like moving average and moving variance on RSSI can be used to detect human presence. They also demonstrated that it is possible to track people's location exploiting the fact that the RSSI patterns of different locations behave differently, and thus can act as a fingerprint to estimate a subject's probable location. Meanwhile, Zhang *et al.* proposed a geometric-model-based method of localization and tracking [3] by relating a link's RSS variance to its line-of-sight (LoS) location relative to the human subjects present. These early works show the possibilities that both the RF variation pattern and the physical model can be used for human tracking and localization. However, as the first step toward RF-based human sensing, these works are still relatively preliminary.

Exposed at the physical layer, channel state information (CSI) provides finer-grained information (with amplitude and phase) than RSSI [4]. With the CSI measurements accessible to the public in 2010 in commodity WiFi chipsets (Intel 5300, Atheros 9580, etc.), the research in WiFi-based device-free human sensing has accelerated. Some RSSI-originated human sensing applications, such as indoor localization [5], are enhanced with CSI information and gain great performance improvement. Many other human behavior recognition applications, which are hard to differentiate using RSSI, also benefit from the capability of fine-grained CSI, including gesture control [6], gait identification [7], fall detection [8, 13], tracking [9], activity recognition [10, 11], vital signs monitoring [12, 14], and so on. From a technical perspective, research efforts have been devoted not only to feature engineering and pattern classification (pattern-based approach) [5, 6, 8, 10, 12, 13] but also to modeling the relationship between signal space and human activity space [7, 8, 11, 14, 15] (model-based approach) to achieve more fine-grained human behavior sensing using WiFi signals. The above works can be categorized according to two dimensions: the problem domain and the solution domain, as shown in Fig. 1. While most of the works fall into pattern-based or model-based approaches, some [7, 11] use the combination of the two approaches as the solution.

In this article, we argue that while pattern-based approaches are intuitive and straightforward for coarse-grained sensing applications, more complex and fine-grained human behavior recognition requires a more general RF model to accurately characterize the relationship between human motion and the resultant signal variations. In this regard, we first introduce the Fresnel zone model for indoor human sensing, and would like to show the superiority of Fresnel-zone-model-based human sensing over pattern-based approaches. We argue that Fresnel-zone-model-based approaches have obvious advantages

The authors first classify the existing work into two categories: pattern-based and model-based recognition solutions. Then they review and examine the two approaches together with their enabled applications. Finally, they show the favorable properties of model-based approaches by comparing them using human respiration detection as a case study.

[1] Assume that human subjects will always carry the device on which the sensing task is performed.

The authors are with the Key Laboratory of High Confidence Software Technologies, Peking University. D. Zhang is the corresponding author.

Digital Object Identifier: 10.1109/MCOM.2017.1700143

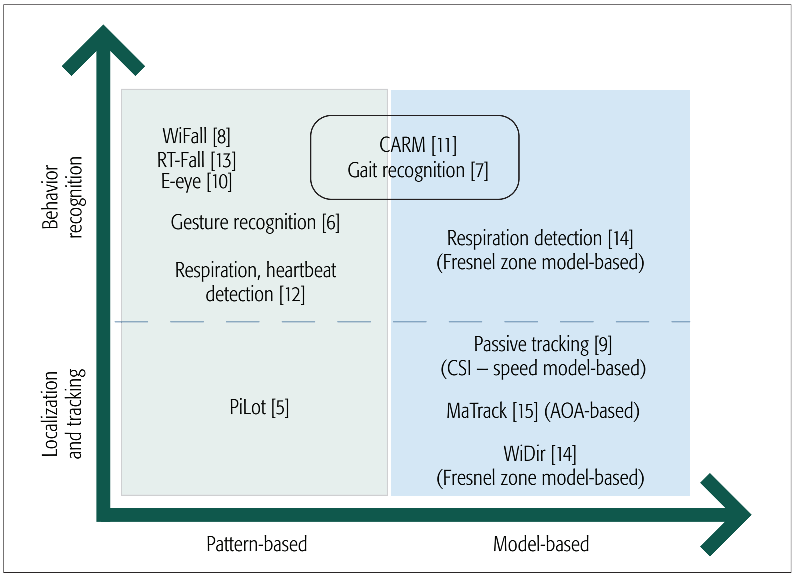


Figure 1. Design space of existing works: problem domain vs. solution domain.

and great potential in achieving centimeter and even millimeter scale human activity sensing [14], enabling a wide spectrum of applications.

Pattern-Based Approaches

Generally speaking, human sensing techniques aim to detect one's rich context information, including presence, location, moving trajectory, activity, gesture, identity, vital sign, interaction with objects, and so on. The goal of RF-based human sensing is behavior recognition based on the radio signals collected on the RF receiver.

To build a human behavior sensing system using radio signals, the connection between signal variations and human activities must be established. If the signal variation patterns have unique and consistent relations with certain human activities, it is possible for a pattern-based (or learning-based) method to recognize human behaviors accurately from signal patterns.

Feature Selection

The key to designing pattern-based approaches is to observe and find discriminative patterns to construct features and differentiate different human behaviors of interest. The features can be very simple or sophisticated, depending on the complexity of the recognition task and the required granularity.

For simple sensing tasks, feature selection is often based on intuition or direct observation. When the number of behaviors that need to be distinguished is small, it is often easy to find regular but differentiable signal patterns. In this case, one or two features may be enough to distinguish among behaviors. For example, for motion detection [2], we observe that any motion causes signal fluctuations. Features such as moving average and moving variance are good indicators; these simple features are often enough for presence sensing purposes.

As both the number of human behaviors and the sensing granularity increase, it becomes challenging to find one-to-one mappings between behaviors and signal patterns. One or two simple features are not enough for this task anymore. In this case, more features are needed to increase the dimension of feature space. For example, WiFall uses seven features for fall detection [8], and WifiU employs a total of 170 features for gait recognition [7]. Meanwhile, a simple threshold-based method may no longer work, so more powerful non-linear classifiers are needed. Often, advanced techniques such as Dynamic Time Wrapper (DTW) are applied to increase the robustness of classification, as shown in [6, 10]. Signal statistic characteristics, such as normalized standard deviation, median absolute deviation, and amplitude histogram of the CSI waveform, are the most common candidate features [8, 10]. More sophisticated features could be obtained with the help of physical models [7, 11]. These features are then fed into general-purpose classifiers such as support vector machines (SVMs) [8] or specially designed classifiers for classification. For pattern-based algorithms with big numbers of features, people began to lose understanding of the relationship between signal features and human behaviors. Therefore, designing a human behavior sensing system requires many rounds of feature adjustments in the feature selection process, which is often labor-intensive but with bounded classification accuracy.

Pattern Consistency

Pattern-based human sensing approaches rely on consistent and differentiable signal patterns for behavior recognition. This means the same patterns are always expected for a specific behavior. If the signal patterns are inconsistent for the same behavior(e.g., the same activity performed at several different locations), the sensing system may face severe performance problems. Different from wearable sensors that are attached to human bodies in fixed positions, contact-free passive sensing with WiFi signals does not assume a fixed position for a human body with respect to WiFi devices. As a result, even the same activity causes very different signal patterns at different locations. Besides the fact that the distance between a subject and WiFi devices affects the signal amplitude received at the receiver, the subject may interact with multipath differently at each location by blocking different path components. Even worse, the orientation of the subject also matters [3]. For example, in E-eye [10] the activity profiles are requested to be associated with a few locations in the home, so activities performed at different locations might not be well recognized.

The converse problem exists in the case when several behaviors share similar signal patterns. This is becoming common in today's sensing tasks, which require fine-grained sensing capabilities. For example, in gait recognition [7], conventional statistical indicators cannot be used as features because the values are almost identical for subtle movements. More discriminative features such as gait cycle length, estimated footstep length, the maximum, minimum, average, and variance for torso and leg speeds during the gait cycle, as well as spectrogram signatures are extracted from the time-frequency domain with the help of models.

In the above two cases, the pattern-based approaches face limitations. However, if certain features can be found through accurate models — for instance, the histogram distribution feature used in E-eye [10] is replaced by the CSI frequency feature in CARM [11] with the help of the CSI-Speed model — the performance of activity recognition is less affected by the subject's relative location. Please note that this is not an easy job using pure empirical observation.

SCALABILITY

The performance of pattern-based approaches relies on the data samples trained and tested. Often, a pattern-based sensing system is built on top of the model learned from a small training dataset of a few people collected at a few locations, and it is thus difficult to scale, suffering performance degradation when deployed in rooms with different sizes or layouts, or changing the positioning of each WiFi device.

Despite the drawbacks in feature selection, environmental dependence, and scalability, pattern-based approaches have been very popular and successful in device-free human behavior sensing applications because they are not only conceptually intuitive but also relatively simple to design, for both data collection and algorithm development.

MODEL-BASED APPROACHES

Different from pattern-based approaches, which often involve nontrivial training effort and could only recognize a limited set of pre-defined activities, model-based approaches are based on the understanding and abstraction of a mathematical relationship among human locations and/or behaviors, the received signals, and the surrounding environment. In the case of device-free human sensing with WiFi CSI, the aim of modeling is to relate the signal space to the physical space including human and environment, and reveal the physical law characterizing the mathematical relationship between the received CSI signals and the sensing target.

MODELS IN THE WILD

Compared to the study of pattern-based approaches, there has been much less research on model-based device-free human behavior sensing with WiFi devices. In this section, we first briefly present the few model-based human sensing research works that have appeared in recent years, and then introduce our proposed Fresnel zone model and its applications in human sensing.

CSI-Speed Model: Wang *et al.* proposed the CSI-speed model, which quantifies the correlation between CSI power (amplitude) dynamics and the speed of path length change of the reflected paths caused by human movement [11]. They find that the total CFR power is the sum of a constant offset and a set of sinusoids, where the frequencies of the sinusoids are functions of the speeds of path length changes.

The importance of the CSI-speed model lies in the fact that it mathematically links the CSI with the speed of reflected path length change due to human body movement. In such a way, the path length change rate information can be extracted from CSI power amplitude by methods like short time Fourier transform (STFT). However, there is no mathematical mapping from the path length change rate to the human motion speed and human activities. As a consequence, approximated speed information is used as input to a pattern-based learning algorithm for behavior recognition. In the CSI-speed-model-based activity recognition system CARM, Wang *et al.* assume that the human motion is half the path length change speed. Although this approximation is not very accurate, for a total of eight predefined daily activities, CARM can differentiate and recognize them well [11].

Based on the CSI-speed model, *Widar* by Qian *et al.* attempts to build a CSI-Mobility model that quantifies the relationship between CSI dynamics and a user's location and velocity for precise tracking [9]. The CSI-Mobility model tries to fill the gap between the path length change rate and the human moving velocity. As the CSI-speed model provides no direction information, the CSI-mobility model estimates the velocity by formulating it into an optimization problem. With the extended model capability, *Widar* is capable of tracking a human's walking direction and velocity. However, the lack of initial position prevents the precise mapping from speed to velocity, hindering accurate trajectory tracking.

Angle of Arrival Model: Angle of arrival (AoA) measurement is a method of determining the direction of propagation of an RF wave incident on an antenna array. AoA can be estimated by the phase difference pattern across antennas of the array. The resolution of AoA grows with the number of antennas. Normally, five to eight antennas are required for a good AoA estimation. Recently, subspace-based methods such as the MUSIC algorithm have been adopted to obtain finer angle estimation. With two or more AoA measurements from known points, the location of the signal source can be computed by triangulation.

In device-free WiFi sensing, the received signal via different reflected paths off a moving person can be viewed coming from one virtual source with the same angle. For a person to be successfully located using the AoA method, the target's angles to two RF receivers should be obtained. Li *et al.* proposed a device-free localization system, *MaTrack* [15]. The rationale for obtaining the AoA of a moving target is that the signals reflected from it keep changing in angle and time delay, which are incoherent with the reflected signals from environmental static objects. Although MaTrack can be used to infer the AoA of a moving target, its angle resolution is not fine enough to separate the reflected paths of the human body, which limits its application in human sensing tasks other than localization.

Fresnel Zone Model: The Fresnel zone concept originated from Augustin Fresnel's research on light's interference and diffraction in the early 19th century. When applied in a radio propagation area, Fresnel zones refer to the series of concentric ellipsoids with two foci corresponding to the transmitter and receiver antennas. Radio waves traveling through the first Fresnel zone are all in-phase, enhancing the signal strength received at the receiver. Successive Fresnel zones alternately provide destructive and constructive interference to the received signal strength at the receiver side [14].

Different from previous work that applies the Fresnel-Kirchhoff knife-edge diffraction model for

Radio waves traveling through the first Fresnel zone are all in-phase, enhancing the signal strength received at the receiver. Successive Fresnel zones alternately provide destructive and constructive interference to the received signal strength at the receiver side.

human sensing, our Fresnel zone model expands the sensing range to the vast regions outside of the first Fresnel zone. In the device-free passive sensing scenario, a pair of WiFi transceivers are placed at the fixed location. When an object appears in the Fresnel zones in free space, the radio signals can be viewed as traveling from the transmitter to the receiver via two paths: one that goes directly (the LoS path) and another that is reflected by the object (the reflected path). The two signals combine to create a superimposed signal at the receiver side. When the object moves, while the signal traveling via the LoS remains the same, the signal reflected by the object changes over time. As the length of the reflected path changes, the relative phase difference between the LoS signal and the reflected signal changes accordingly, and the received signal will present peaks or valleys when the object crosses the boundaries of the Fresnel zones. The situation is similar in a real multipath-rich environment. In this case the Fresnel zone model can be approximated in such a way that the LoS signal is superimposed with the multiple reflected signals from the environmental static objects, and the signals reflected from the moving object are unified and simplified into one dominant signal component that changes over time. Mathematically, the Fresnel zone model characterizes the relationship between the geometrical position of the sensing target and the induced CSI power amplitude variations caused by the motion of the target.

> The power of the Fresnel zone model is that it not only reveals the relationship between the centimeter-scale or even millimeter-scale human activities and the received WiFi signals, but also describes how the received signals vary for a human activity performed at different locations and orientations.

The power of the Fresnel zone model is that it not only reveals the relationship between the centimeter-scale or even millimeter-scale human activities and the received WiFi signals, but also describes how the received signals vary for a human activity performed at different locations and orientations [14]. This capability makes the Fresnel zone model location-aware, different from the CSI-speed or CSI-Mobility model.

In order to show how the Fresnel zone model can be used for human behavior recognition, we leverage the properties of multiple subcarriers in the WiFi received signals and build the human indoor walking direction and distance estimation system *WiDir* [14].

For multiple subcarriers with different wavelengths in commodity WiFi devices, their corresponding Fresnel zones are of slightly different sizes. As a consequence of a person moving inward/outward, it would cross the Fresnel zone boundaries of different subcarriers in sequence and generate increasing/decreasing time delays between a fixed pair of subcarriers. Then the inward/outward walking direction can be determined by inspecting the CSI time delay between two WiFi subcarriers. Furthermore, as crossing Fresnel zone boundaries corresponds to a series of peaks and valleys in the CSI waveform, WiDir counts the peaks and valleys in each axis for distance estimation. With two pairs of WiFi devices, both direction and distance in the 2D plane can be estimated directly and accurately. The WiDir example showcases that sensing the indoor human walking direction can be achieved leveraging only the Fresnel zone model. It is different from the CSI-speed or CSI-mobility model for human behavior recognition, where only the reflected path change rate is extracted from the model, while the human speed or human activities are approximately obtained using the path change rate and other information.

Discussion

Through the introduction of the above three lines of model-based human sensing research, it can be seen that model-based approaches have the advantage of leveraging physical laws and having clear physical interpretations. Hence, we could use the derived models to accurately extract certain parameters from the received signals and solve a class of problems. For example, the CSI-speed model can precisely sense the speed information, which can support applications such as activity recognition, gait recognition, and tracking. The AoA model is geometry-related, which suits localization applications. While the above models generally target at obtaining a specific output such as speed, velocity or angle, Fresnel Zone model seems to be more general as the basis for understanding how human motion affects the received RF signal and further designing various human behavior recognition systems, as can be seen in the walking direction sensing application as well as the human respiration detection application, which is presented in the next section.

Case Study: Respiration Sensing

In order to demonstrate the generality and potential of our proposed Fresnel zone model in human behavior recognition, in this section we use human respiration detection as an application example to show how the existing pattern-based approaches and our proposed solution achieve the goal. We further compare the advantages and drawbacks of these approaches, and argue that our proposed Fresnel zone model is not only general in supporting a wide spectrum of applications, but also very powerful in revealing the sensing limit as well as the complex relationship among human motion/location/orientation, the received CSI of different subcarriers, as well as the physical environment including WiFi devices.

Human respiration detection using commodity WiFi CSI has been explored in recent years. In [12], by observing the obvious periodic sinusoid-like patterns that appear in the received WiFi CSI across different subcarriers, which seemed to have a high correlation with human respiration, Liu *et al.* developed a WiFi CSI pattern-based vital sign monitoring system. In this work, it is assumed that the sinusoid-like pattern exists in at least one of the subcarriers; they focus on proposing methods for signal processing and respiration rate extraction, which include the steps of filtering, peak-to-peak time interval measurement, and power spectral density (PSD)-based K-means clustering. To ensure that the appropriate subcarriers are selected, a variance with a predefined threshold is employed before processing [12].

With the Fresnel zone model, we reexamined the same human respiration sensing problem [14]. According to the WiFi signal propagation properties in the Fresnel zones, when an object crosses a series of Fresnel zones, the received signal shows a continuous sinusoidal-like wave. If a moving object causes a reflected signal path length change shorter than a wavelength (e.g.,

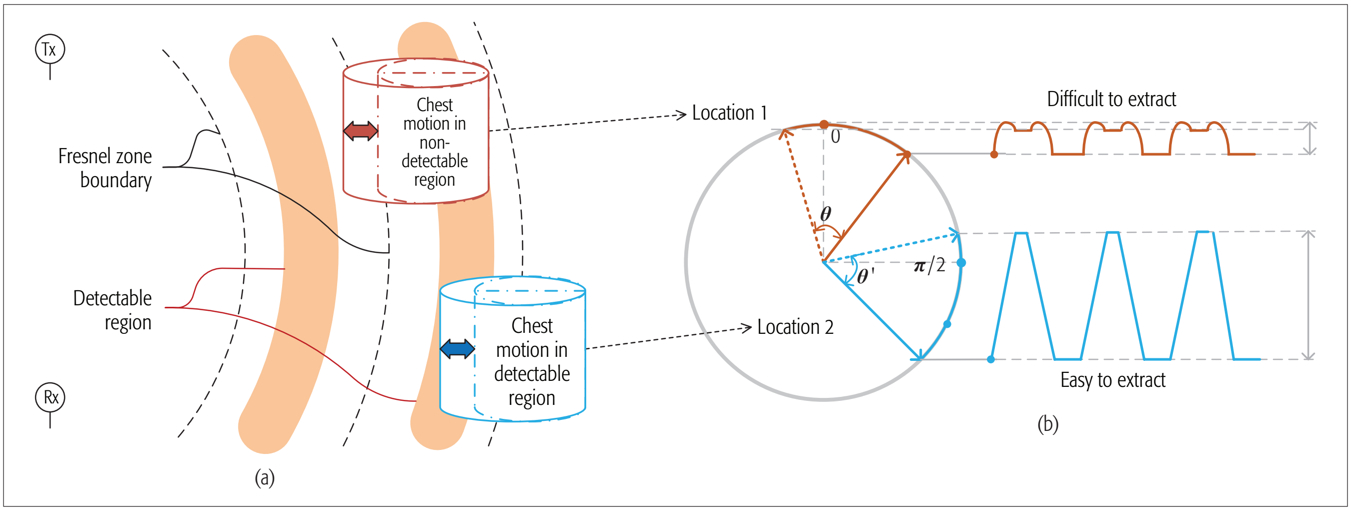


Figure 2. a) Human respiration detection at two locations in Fresnel zones; b) their corresponding CSI waveforms.

5.7 cm for 5.24 GHz), the received signal is just a fragment of the sinusoidal cycle depending on the location of the object in the Fresnel zones. As the human chest motion displacement due to respiration is around 5 mm and the resultant reflected path length change is far less than a wavelength, it roughly corresponds to a phase change of 60° in one sinusoidal cycle [14]. Given the phase change in a sinusoidal cycle (as shown in Fig. 2b), both the angle of phase change and its position affects the shape of the resultant signal waveform. Apparently, in order to make the respiration rate easy to extract correctly, it is expected that the angle not only covers a large range but also lies fully in the monotonically changing fragment of the cosine wave. Based on the above study, Zhang *et al.* conclude that within each Fresnel zone, the worst human location for centimeter-level motion sensing is around the boundary, while the best location appears in the middle of the Fresnel zone, as illustrated in Fig. 2. By further considering the multi-frequency diversity of subcarriers, a respiration detection map can be constructed to instruct where respiration is detectable by different subcarriers, as illustrated in Fig. 3. According to this map, Zhang *et al.* found that in the inner Fresnel zones, there are many places where human chest movement cannot be detected by any subcarrier, while a short human body move outwards would make the human respiration detectable; in the frequency diversity-enabled region, at least one subcarrier can be used to detect the human respiration according to the ideal Fresnel zone model. Besides the impact of location, they also show that the orientation of the human body also matters. Different orientations lead to different effective chest displacements with respect to a pair of transceivers, thus influencing the detectability of human respiration [14].

To validate our observation, we conducted extensive experiments in an apartment. The experiment settings are illustrated in Fig. 3b. A subject laid on the bed facing up. In order to validate that the detectable and undetectable regions alternatively appeared in the shape of ellipses in the geometrical space, we mounted a COTS WiFi transmitter and receiver pair on two vertically placed slide rails. We examined six consecutive Fresnel zones and collected 2 hours' CSI data at a sampling rate of 20 packets/s. The results show that the detectable and undetectable regions indeed appear alternatively by fixing the human posture and moving the LoS away from the human continuously, and the estimation performance in the detectable regions is consistent. In our case, the median estimation errors of respiration rate in the three detectable regions are about 0.09 breaths per minute (bpm), 0.15 bpm, and 0.06 bpm, respectively, compared to the overall mean estimation error of 0.4 bpm reported in [12]. Please note that our experiment results show that human respiration cannot be monitored reliably in the three undetectable regions, which were not reported in the previous work.

By comparing the above two human respiration sensing approaches, it can be seen that the pattern-based respiration detection method used in [12] is intuitive and works well as long as the assumption holds, that is, at least one subcarrier is able to sense the human chest movement. However, our proposed Fresnel-zone-model-based approach could explain when the pattern-based system works or not, why some locations and subcarriers are not able to detect human respiration effectively, and how WiFi devices should be positioned for better respiration monitoring [14]. With these findings and understandings, designing a practical respiration monitoring system should consider many factors such as the location of the subject, the posture of the subject, and the positioning of WiFi devices for effective continuous monitoring. These considerations also apply to situations where more than one person's respiration rates are monitored.

From the above case study we can see that not only can the Fresnel zone model interpret where and at what orientation a person's respiration can be sensed, but it can also guide the sensing system design. However, the pattern-based respiration sensing approach can only sense respiration when there are obvious and clear signal patterns. It can neither answer the question why sometimes human respiration cannot be sensed, nor provide guidance on how to design a robust monitoring system.

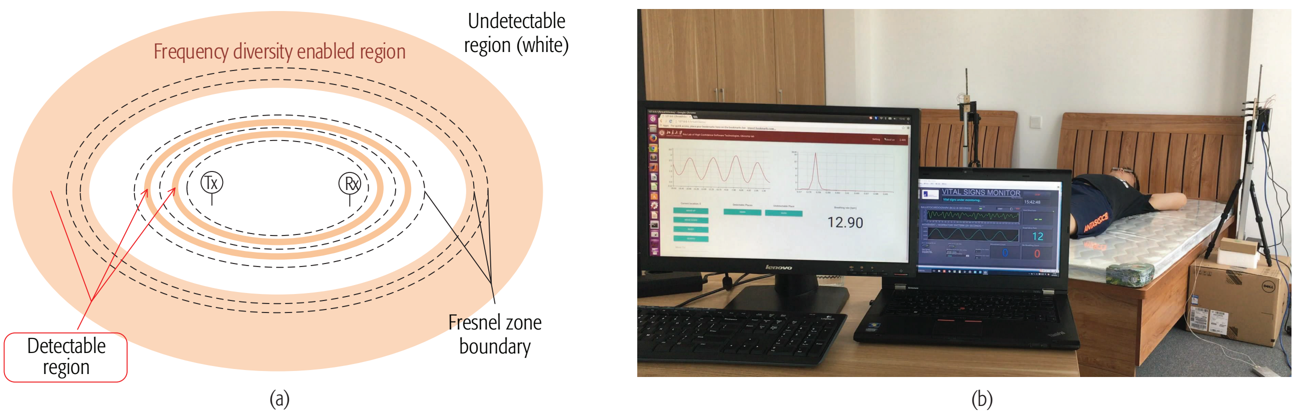

Frequency diversity enabled region
Undetectable region (white)
Tx
Rx
Detectable region
Fresnel zone boundary
(a)
(b)

Figure 3. a) Respiration detection map for multi-subcarrier Fresnel zones; b) experiment settings for respiration detection.

Conclusion

Today's device-free WiFi CSI-based human behavior recognition works are either pattern-based or model-based, or a combination of both. While the pattern-based approaches are straightforward and effective for many sensing applications, they are observation- and empirical-study-based, and usually require a deep case-by-case investigation and intensive training for a specific application. Although they have been very popular and successful in this field, pattern-based approaches are trial and error by nature, having the bottleneck of predicting the sensing limit and understanding what complexity of human behaviors is recognizable, especially for continuous and fine-grained human sensing tasks.

Model-based approaches aim to fundamentally understand the governing law on how a human's motion/location/orientation impacts the received signals in the environments, and mathematically depict the direct relationship between the received signals and the sensing target. For this reason, model-based approaches not only have the potential to solve more complex and fine-grained human behavior recognition problems, but also could guide us in understanding the sensing limits (e.g., sensing area, fineness of behavior, accuracy bound) and the rationale behind it in the real world as well. Among all the efforts, the Fresnel zone model seems to be the most general one. It not only shows its effectiveness in supporting both coarse-grained and fine-grained human behavior recognition applications, but also helps us to understand how radio waves propagate in real-world environments and what is the possible sensing limit with WiFi CSI measurements. With those attributes, we believe that the Fresnel zone model has the potential to revolutionize the RF-based human sensing field and enable more real-world applications, which were not possible without the model.

However, there is still a lot of research that needs to be done in order to fully understand the properties of the Fresnel zone model in the multipath-rich indoor environments, especially with multiple moving objects. It also should be noted that there is no single model which can solve all the problems. With those points in mind, while we strongly encourage researchers in the WiFi human sensing field to join us in developing new models and improving the existing models due to their obvious advantages, we envision that combining the model-based approaches with the pattern-based approaches would still be the most effective way for WiFi CSI-based human behavior recognition in the coming years.

Acknowledgment

This research is supported by the National Key Research and Development Plan under Grant No. 2016YFB1001200, and Peking University Information Technology Institute (Tianjin Binhai).

References

[1] K. Woyach, D. Puccinelli, and M. Haenggi, "Sensorless Sensing in Wireless Networks: Implementation and Measurements," *Proc. Int'l. Symp. Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, 2006, pp. 1–8.
[2] M. Youssef, M. Mah, and A. Agrawala, "Challenges: Device-Free Passive Localization for Wireless Environments," *Proc. ACM Int'l. Conf. Mobile Computing and Networking*, 2007, pp. 222–29.
[3] D. Zhang *et al.*, "An RF-Based System for Tracking Transceiver-Free Objects," *Proc. IEEE Int'l. Conf. Pervasive Computing and Commun.*, 2007, pp. 135–44.
[4] Z. Yang, Z. Zhou, and Y. Liu, "From RSSI to CSI: Indoor Localization Via Channel Response," *ACM Computing Surveys*, vol. 46, no. 2, 2013, p. 25.
[5] J. Xiao *et al.*, "Pilot: Passive Device-Free Indoor Localization Using Channel State Information," *Proc. IEEE Int'l. Conf. Distributed Computing Systems*, 2013, pp. 236–45.
[6] P. Melgarejo *et al.*, "Leveraging Directional Antenna Capabilities for Fine-Grained Gesture Recognition," *Proc. ACM Int'l. Joint Conf. Pervasive and Ubiquitous Computing*, 2014, pp. 541–51.
[7] W. Wang, A. X. Liu, and M. Shahzad, "Gait Recognition Using WiFi Signals," *Proc. ACM Int'l. Joint Conf. Pervasive and Ubiquitous Computing*, 2016, pp. 363–73.
[8] C. Han *et al.*, "WiFall: Device-Free Fall Detection by Wireless Networks," *Proc. IEEE Conf. Computer Commun.*, 2014, pp. 271–79.
[9] K. Qian *et al.*, "Decimeter Level Passive Tracking with WiFi," *Proc. ACM Wksp. Hot Topics in Wireless*, 2016, pp. 44–48.
[10] Y. Wang *et al.*, "E-Eyes: Device-Free Location-Oriented Activity Identification Using Fine-Grained Wifi Signatures," *Proc. 20th ACM Int'l. Conf. Mobile Computing and Networking*, 2014, pp. 617–28.
[11] W. Wang *et al.*, "Understanding and Modeling of WiFi Signal Based Human Activity Recognition," *Proc. 21st ACM Int'l. Conf. Mobile Computing and Networking*, 2015, pp. 65–76.
[12] J. Liu *et al.*, "Tracking Vital Signs During Sleep Leveraging Off-the-Shelf WiFi," *Proc. ACM Int'l. Symp. Mobile Ad Hoc Networking and Computing*, 2015, pp. 267–76.
[13] H. Wang *et al.*, "RT-Fall: A Real-Time and Contactless Fall Detection System with Commodity WiFi Devices," *IEEE Trans. Mobile Computing*, vol. 16, no. 2, 2017, pp. 511–26.
[14] D. Zhang, H. Wang, and D. Wu, "Toward Centimeter-Scale Human Activity Sensing with Wi-Fi Signals," *IEEE Computer*, vol. 50, no. 1, 2017, pp. 48–57.

[15] X. Li *et al.*, "Dynamic-Music: Accurate Device-Free Indoor Localization," *Proc. ACM Int'l. Joint Conf. Pervasive and Ubiquitous Computing*, 2016, pp. 196–207.

Biographies

Dan Wu (dan@pku.edu.cn) is a Ph.D. student in computer science in the School of Electronics Engineering and Computer Science at Peking University, China. His research interests include software modeling, mobile sensing, and ubiquitous computing. He received a B.S. in computer science from the University of Science and Technology of Beijing.

Daqing Zhang (dqzhang@sei.pku.edu.cn) is a chair professor with the Key Laboratory of High Confidence Software Technologies, Peking University. He obtained his Ph.D. from the University of Rome "La Sapienza," Italy, in 1996. His research interests include context-aware computing, urban computing, mobile computing, big data analytics, and more. He is an Associate Editor for *ACM Transactions on Intelligent Systems and Technology*, *IEEE Transactions on Big Data*, and other publications.

Chenren Xu (chenren@pku.edu.cn) is an assistant professor in the School of Electronics Engineering and Computer Science at Peking University. He received a B.E. in automation from Shanghai University in 2008, an M.S. in applied mathematical statistics, and a Ph.D. in electrical and computer engineering from Rutgers University in 2014. His research interests span wireless communication to network-centric application from a computer system perspective.

Hao Wang (wanghao@sei.pku.edu.cn) received his Ph.D. degree in computer science from Peking University in 2017. He is going to be a research engineer for Huawei Technologies Co. Ltd. His research interests include mobile crowdsensing, ubiquitous computing, and smart environment.

Xiang Li (lixiang13@pku.edu.cn) is a Ph.D student in computer science in the School of Electronics Engineering and Computer Science at Peking University. His research interests include mobile computing, mobile sensing, and ubiquitous computing. He received a B.S. in computer science from the School of Electronics Engineering and Computer Science at Peking University.

A Survey on Behavior Recognition Using WiFi Channel State Information

Siamak Yousefi, Hirokazu Narui, Sankalp Dayal, Stefano Ermon, and Shahrokh Valaee

The authors present a survey of recent advances in passive human behavior recognition in indoor areas using the CSI of commercial WiFi systems. The movement of the human body parts cause changes in the wireless signal reflections, which result in variations in the CSI. By analyzing the data streams of CSI for different activities and comparing them against stored models, human behavior can be recognized.

The work is supported in part by the Fund de Researche du Quebec-Nature et technologies (FQRNT) and FURUKAWA ELECTRIC Co., LTD.

The authors would like to thank Dr. Wei Wang of Nanjing University, China for the helpful discussions on CARM and sharing part of his code.

Digital Object Identifier:
10.1109/MCOM.2017.1700082

ABSTRACT

In this article, we present a survey of recent advances in passive human behavior recognition in indoor areas using the channel state information (CSI) of commercial WiFi systems. The movement of the human body parts cause changes in the wireless signal reflections, which result in variations in the CSI. By analyzing the data streams of CSIs for different activities and comparing them against stored models, human behavior can be recognized. This is done by extracting features from CSI data streams and using machine learning techniques to build models and classifiers. The techniques from the literature that are presented herein have great performance; however, instead of the machine learning techniques employed in these works, we propose to use deep learning techniques such as long-short term memory (LSTM) recurrent neural networking (RNN) and show the improved performance. We also discuss different challenges such as environment change, frame rate selection, and the multi-user scenario; and finally suggest possible directions for future work.

BACKGROUND ON TRADITIONAL ACTIVITY RECOGNITION SYSTEMS

Human activity recognition has gained tremendous attention in recent years due to numerous applications that aim to monitor the movement and behavior of humans in indoor areas. Applications include health monitoring and fall detection for elderly people [1], contextual awareness, activity recognition for energy efficiency in smart homes [2], and many other Internet of Things (IoT)-based applications [3].

In existing systems, the individual has to wear a device equipped with motion sensors such as a gyroscope and an accelerometer. The sensor data is processed locally on the wearable device or transmitted to a server for feature extraction, and then supervised learning algorithms are used for classification. This type of monitoring is known as *active* monitoring. The performance of such a system is shown to be around 90 percent for recognition of activities such as sleeping, sitting, standing, walking, and running [4].

However, always wearing a device is cumbersome and may not be possible for many passive activity recognition applications, where the person may not be carrying any sensor or wireless device. While camera-based systems can be used for passive activity recognition, the line-of-sight (LOS) requirement is a major limitation for such systems. Furthermore, the camera-based approaches have privacy issues and cannot be employed in many environments. Therefore, a passive monitoring system based on wireless signal, which does not violate the privacy of people, is desired.

Because of ubiquitous availability in indoor areas, recently, WiFi has been the focus of much research on activity recognition. Such systems consist of a WiFi access point (AP) and one or several WiFi enabled device(s) located in different parts of the environment. When a person engages in an activity, body movement affects the wireless signals and changes the multi-path profile of the system.

TECHNIQUES BASED ON WI-FI SIGNAL POWER

Received signal strength (RSS) has been used successfully for active localization of wireless devices using WiFi fingerprinting techniques as summarized in [5]. RSS has also been used as a metric for passive tracking of mobile objects [6]. When a person is located between a WiFi device and an AP, the signal is attenuated, and hence a different RSS is observed. Although RSS is very simple to use and can easily be measured, it cannot capture the real changes in the signal due to the movement of the person. This is because RSS is not a stable metric even when there is no dynamic change in the environment [7].

TECHNIQUES REQUIRING MODIFIED WIFI HARDWARE

To use some metrics other than RSS, in some systems, the WiFi system is modified so that extra information can be extracted from the signal. The WiFi universal software radio peripheral (USRP) software radio system is a modified WiFi hardware and has been used for 3D passive tracking in WiTrack [7]. The idea is to measure the Doppler shift in the orthogonal frequency-division multiplexing (OFDM) signals caused by movement of the human body using a technique called frequency modulated carrier wave (FMCW). Since the Doppler shift is related to the distance, the location of the target can be estimated. Using a similar idea to WiTrack, in WiSee [2], the USRP system is used to measure the Doppler shift in OFDM signals due to movement of the human body. The movement of the parts of the body toward the receiver causes positive Doppler shift, while moving the body parts away results in negative shift. For instance, for a gesture moving at 0.5m/s, in a 5 GHz system, the Doppler shift is

Siamak Yousefi and Shahrokh Valaee are with the University of Toronto; Hirokazu Narui, Sankalp Dayal, and Stefano Ermon are with Stanford University.

0163-6804/17/$25.00 © 2017 IEEE

around 17 Hz [2]. Therefore, such small Doppler shifts need to be detected in the system. In WiSee, the received signal is transformed into narrowband pulses of a few Hertz, and the WiSee receiver tracks the Doppler shift in the frequency of these pulses. After transforming the wideband 802.11 to narrowband pulses, the next steps in WiSee are as follows.

Doppler Extraction: To extract the Doppler information, WiSee computes the frequency-time Doppler profile by taking the fast Fourier transform (FFT) over samples in a window of half a second and then shifting the window by 5 ms and continuing this process. This technique is also known as short-time Fourier transform (STFT), which was used in other techniques as well [8, 9]. Since the movement of a human body generally has a speed of 0.25 m/s to 4 m/s, the Doppler shifts at 5 GHz is between 8 Hz and 134 Hz, hence only the FFT output in this frequency range is considered in WiSee.

Segmentation: The next step is to segment the STFT data to distinguish different patterns. For example, a gesture might consist of one segment with positive and negative Doppler shifts, or two or more segments, each of which has a positive and negative Doppler shift. Detecting a segment is based on the energy detection over a small duration. If the energy is 3 dB higher than the noise level, the beginning of the segment is found, and if it is less than 3 dB, the segment has ended.

Classification: The idea of classification is quite simple. Each segment has three possibilities: only positive Doppler shifts, only negative Doppler shift, and segments with both positive and negative shifts, based on which three numbers are assigned to them. Thus, each gesture is represented by a sequence of numbers. The classification task is to compare the obtained sequence with the one used during training.

WiSee also claims that the system can detect multiple moving targets and identify their activities using the idea that the reflections from each mobile target can be regarded as a signal from a wireless transmitter. Therefore, using the idea used in multiple-input multiple-output (MIMO) receivers, the reflected signals due to different people moving in the area can be separated. The problem is to find the weight matrix that, when multiplied with the Doppler energy corresponding to each segment of each antenna, maximizes the Doppler of each segment. To this end, iterative algorithms have been employed.

In contrast to techniques such as WiSee that require specialized USRP software radios, there have been several efforts to employ commercial WiFi APs without the need to modify the WiFi system. To represent the dynamic changes in the environment due to movement of the human body, recently other metrics have been employed, such as channel state information (CSI), which is described in more detail below.

WiFi Channel State Information

CSI of a WiFi System

The wireless devices in IEEE 802.11n/ac standards use MIMO systems. By using MIMO technology, it is possible to increase the diversity gain, array gain, and multiplexing gain, while reducing the co-channel interference [10]. The modulation used in IEEE 802.11 is OFDM where the bandwidth is shared among multiple orthogonal subcarriers. Due to the small bandwidth, the fading that each subcarrier faces is modeled as flat fading. Therefore, using OFDM, the small-scale fading property of the channel can be mitigated.

Let M_T denote the number of transmit antennas at the device, and M_R the number of receive antennas at the AP. The MIMO system at any time instant can be modeled as $\mathbf{y}_i = \mathbf{H}_i\mathbf{x}_i + \mathbf{n}_i$, for $i \in \{1, ..., S\}$ where S is the number of OFDM subcarriers, and $\mathbf{x}_i \in \mathbb{R}^{M_T}$ and $\mathbf{y}_i \in \mathbb{R}^{M_R}$ represent the transmit and received signal vectors for the ith subcarrier, respectively, and $\mathbf{n}_i$ is the noise vector. The channel matrix for the ith subcarrier $\mathbf{H}_i$, which consists of complex values, can be estimated by dividing the output signal with a known sequence of input also known as pilot. The channel matrix is also known as the CSI, as it shows how the input symbol is affected by the channel to reach at the receiver. In OFDM systems, each subcarrier faces a narrowband fading channel, and by obtaining the CSI for each subcarrier, there will be diversity in the observed channel dynamics. This is the main advantage of using CSI compared to RSS, in which the changes are averaged out over all the WiFi bandwidth and hence cannot capture the change at certain frequencies. In some commercial network interface cards (NICs), such as Intel NIC 5300, the CSI can be collected using the tool provided in [11].

In OFDM systems, each subcarrier faces a narrowband fading channel, and by obtaining the CSI for each subcarrier, there will be diversity in the observed channel dynamics.

Limitations and Errors of WiFi Systems

The amplitude of CSI is generally a reliable metric to use for feature extraction and classification, although it can change with transmission power, and transmission rate adaptation. As discussed later, by using filtering techniques, the burst noise can be reduced [9]. However, in contrast to amplitude, the phase of a WiFi system is affected by several sources of error such as carrier frequency offset (CFO) and sampling frequency offset (SFO). The CFO exists due to the difference in central frequencies (lack of synchronization) between the transmitter and receiver clocks. The CFO for a period of 50 μs of 5 GHz WiFi band can be as large as 80 kHz, leading to phase change of 8π. Therefore, the phase changes due to the movement of the body, which is generally smaller than 0.5π, is not observable from CSI phase. The other source of error, SFO, is generated by the receiver analog-to-digital converter (ADC). The SFO is also varying by subcarrier index; therefore, each subcarrier faces a different error.

Due to the unknown CFO and SFO, using the raw phase information may not be useful. However, a linear transformation is proposed in [12], such that the CFO and SFO can be removed from the calibrated phase. This process is also known as phase sanitization. In Fig. 1, the CSI amplitude, CSI phase, and sanitized CSI phase vs. the subcarrier index are plotted for a scenario where the WiFi transmitter and receiver are located in the vicinity of each other in LOS condition. As observed, the CSI amplitude is relatively stable but forms some clusters, as mentioned in [12]. The raw phase increases with subcarrier index since the SFO grows with subcarrier index, as illustrated in Fig. 1b. After phase sanitization, the

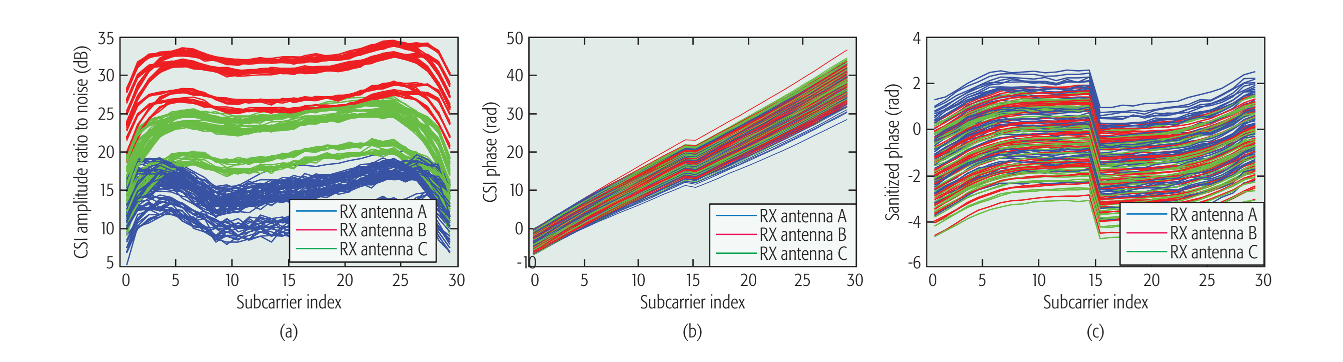


Figure 1. CSI measured in LOS condition for three antennas as a function of subcarrier index: a) amplitude of CSI; b) phase of CSI; c) sanitized phase of CSI.

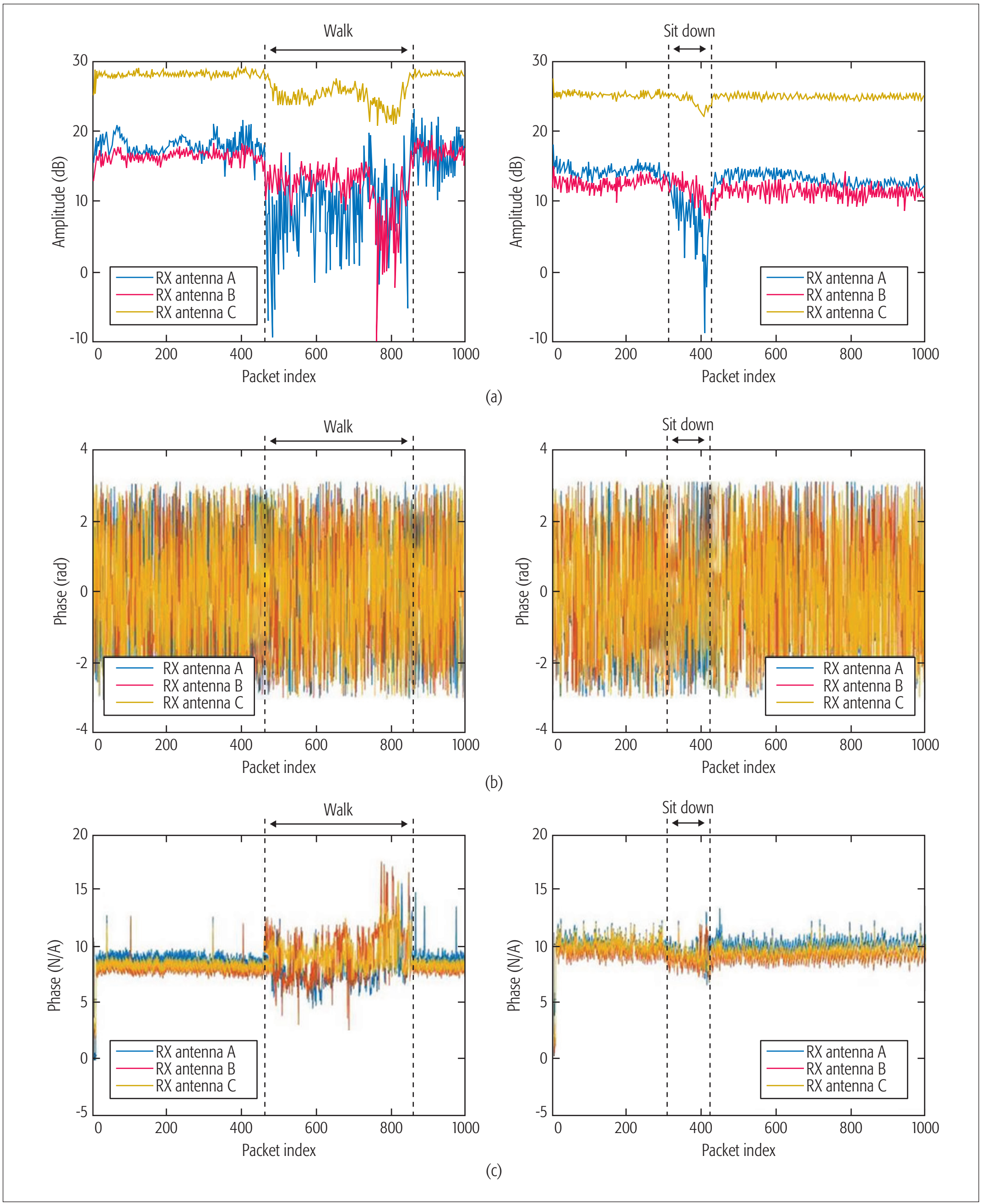


Figure 2. CSI changes under human motion: a) CSI amplitude for three antennas as a function of time; b) CSI phase for three antennas as a function of time; c) sanitized CSI phase for three antennas as a function of time.

change of phase due to SFO will be reduced as observed in Fig. 1c.

Effect of Human Motion on Wireless Channel

The movement of humans and objects change the multipath characteristic of the wireless channel and hence the estimated channel will have a different amplitude and phase. The CSI amplitude for one subcarrier and all the antennas, related to a person walking and sitting down between a WiFi transmitter and receiver, is illustrated in Fig. 2a. The person is stationary for the first 400 packets but then starts walking or sitting down. As observed, when the person is not moving, the CSI amplitudes for all antennas are relatively stable; however, when the activity starts, the CSIs start changing drastically. The walking period is longer than sitting in this experiment because when the person sits down he/she remains stationary.

The received phase, is very distorted due to the CFO and SFO, as mentioned earlier. This can be observed in Fig. 2b. However, using the phase sanitization technique, the effect of errors in phase can be eliminated. The calibrated phase can be observed in Fig. 2c.

Wi-Fi CSI-Based Behavior Recognition

In this section, we provide a summary of the techniques using commercial WiFi NICs. The general diagram of activity recognition systems using WiFi CSI is illustrated in Fig. 3.

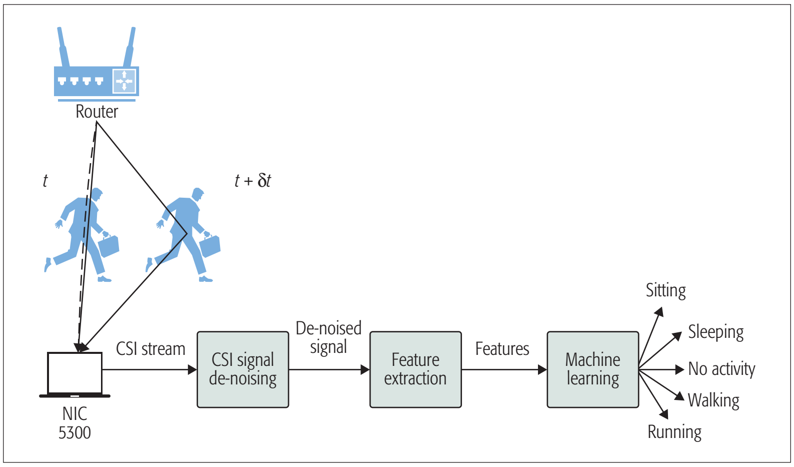


Figure 3. The scheme of common activity recognition techniques. A person is moving in the area between the router and WiFi device from time *t* to time *t* + δ*t*.

Histogram-Based Techniques

One of these technique is E-Eyes [13], in which CSI histograms are used as fingerprints in a database. In the test phase, by comparing the histogram of the obtained CSI with the database, the closest one is found, and hence the activity can be recognized. The preprocessing steps are low-pass filtering and modulation and coding scheme (MCS) index filtering. The former is necessary to remove the high frequency noise, which may not be due to the human movement, and the latter is needed to reduce the unstable wireless channel variations. Although the performance of this technique is good and its computational cost is low, the histogram technique is sensitive to environment changes and hence may not perform well for varying environments.

CARM

Recently, other techniques have been proposed such as WiHear [3], CARM [9[, and the one proposed in [14]. In WiHear, directional antennas are used to capture CSI variations caused by the movement of mouth. The performance of WiHear is good; however, the application is only to monitor spoken words. In [14], the authors use advanced feature extraction and machine learning techniques for recognition of words typed on a keyboard. The idea is similar to the idea in CARM [9], which is described in more detail below.

CSI De-Noising: The CSI is noisy and may not show distinctive features for different activities. Therefore, it is necessary to first filter out the noise and then extract some features to be used for classification using machine learning techniques. There are different methods for filtering the noise such as Butterworth low-pass filters [9]. However, due to the existence of burst and impulse noises in CSI, which have high bandwidths, the low-pass filter cannot yield a smooth CSI stream [9].

It has been shown that there are better techniques for this purpose such as principal component analysis (PCA) de-noising [9]. PCA is a technique for dimensionality reduction of a large-dimension system using the idea that most of the information about the signal is concentrated over some of the features. In CARM, the first principal component is discarded to reduce the noise, and the next five ones are employed for feature extraction. By removing the first principal component, the information due to the dynamic reflection coming from a mobile target is not lost because it is also captured in other principal components. After PCA de-noising of CSI data, some features are extracted from it so that it can be used for classification. Feature extraction is discussed below.

Feature Extraction: One way to extract features from a signal is to transform it to another domain, such as frequency domain. The FFT, which is an efficient implementation of discrete Fourier transform (DFT), can be used for this purpose. To this end, a window size of a certain number of CSI samples is selected, and then the FFT is applied on each segment by sliding the window. This technique, also known as STFT, can detect the frequency changes of a signal over time. The STFT has been applied on radar signals for detection of the movement of torso and legs in [8]. In Fig. 4, the STFT (spectrogram) of CSI for different activities is shown for CSI data collected at 1 kHz rate. As observed in Fig. 4, activities that involve drastic movements such as walking and running show high energy in high frequencies in the spectrogram.

In [3, 9, 14], DWT is employed to extract features from CSI as a function of time. DWT provides high time resolution for activities with high frequencies and high frequency resolution for activities with low speeds. Each level of DWT represents a frequency range, where the lower levels contain higher frequency information while higher levels contain lower frequencies. The advantage of DWT over STFT as mentioned in [9] are:

- DWT can provide a nice trade-off in the time and frequency domains.

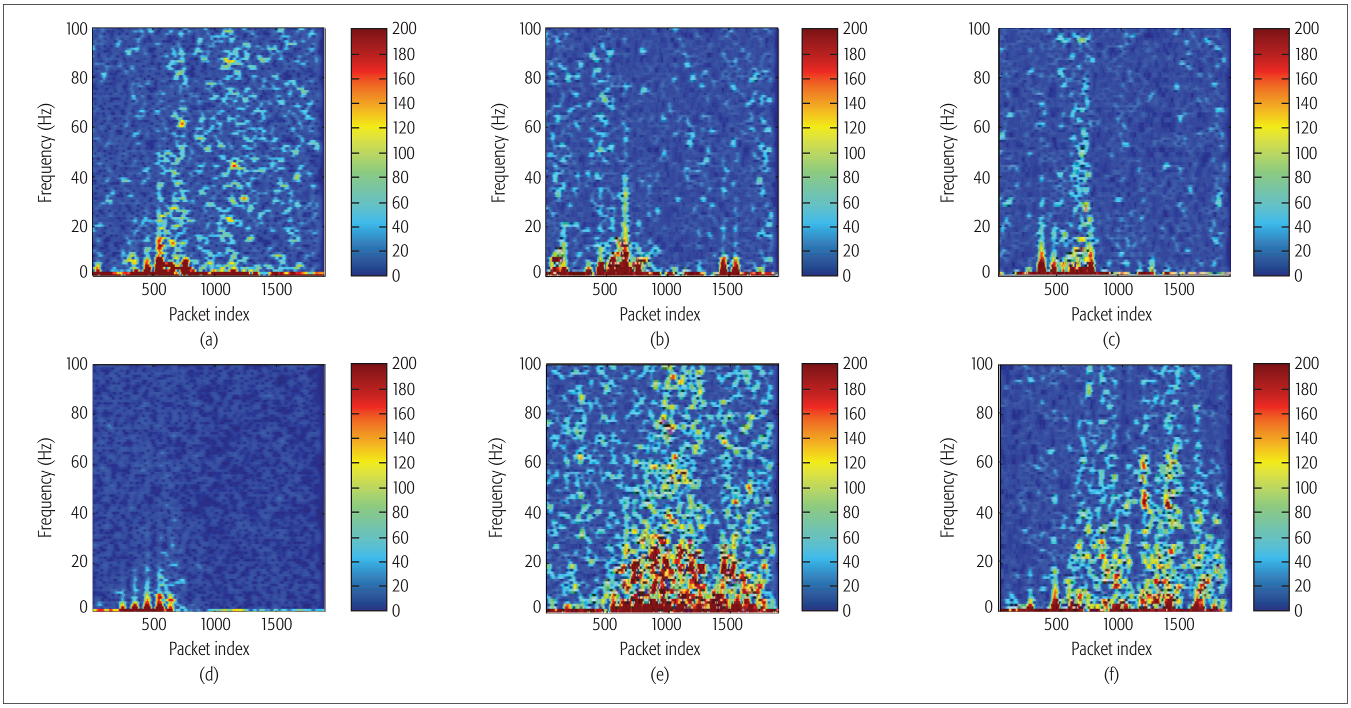


Figure 4. The spectrogram of one subcarrier's CSI amplitude for different activities: a) standing up; b) sitting down; c) lying down; d) falling; e) walking; f) running.

- DWT reduces the size of the data as well, so it becomes suitable for machine learning algorithms.

In CARM, a 12-level DWT is employed to decompose the five principal components (after removing the first principal component). Then the five values of the DWT are averaged. For every 200 ms, CARM extracts a 27-dimensional feature vector including three sets of features:

- The energy in each wavelet level, representing the intensity of movements with different speeds.
- The difference in each level between consecutive 200 ms intervals.
- The torso and leg speeds estimated using the Doppler radar technique [8].

These features are used as the input to the classification algorithm described below.

Machine Learning for Classification: Different machine learning techniques can be used for multi-class classification based on certain features that are extracted. Some of the popular classification techniques are logistic regression, support vector machines (SVMs), hidden Markov model (HMM), and deep learning. Since the activity data is in a sequence, CARM uses HMM, and it is shown that satisfactory results can be obtained.

Using Deep Learning for Behavior Recognition

The problem of activity recognition is somewhat similar to the speech recognition process, where traditionally HMM has been used for classification. However, deep recurrent neural networking (RNN) has been considered as a counterpart of HMM. Training an RNN is difficult as it suffers from the vanishing or exploding gradient problem; however, it was shown in [15] that using the long short-term memory (LSTM) extension of RNN, the best accuracy for speech recognition so far can be achieved. Therefore, we propose using LSTM for activity recognition rather than other conventional machine learning techniques, such as HMM, although feature extraction is not done similar to CARM. Using LSTM has two advantages. First, the LSTM can extract the features automatically; in other words, there is no necessity to pre-process the data. Second, LSTM can hold temporal state information of the activity, i.e., LSTM has the potential to distinguish similar activities like "Lie down" and "Fall." Since "Lie down" consists of "Sit down" and "Fall," the memory of LSTM can help in recognition of these activities.

Evaluation of Different Methods

In this section, we implement different methods as well as our proposed method and show the performance of each one.

Measurement Setup

We do the experiments in an indoor office area where the Tx and Rx are located 3 m apart in LOS condition. The Rx is equipped with a commercial Intel 5300 NIC, with sampling rate of 1 kHz. A person starts moving and doing an activity within a period of 20 s in LOS condition, while in the beginning and at the end of the time period the person remains stationary. We also record videos of activities so that we can label the data. Our dataset includes 6 persons, 6 activities, denoted as "Lie down, Fall, Walk, Run, Sit down, Stand up," and 20 trials for each one.

Evaluating Machine Learning Techniques

We apply the PCA on the CSI amplitude, and then use STFT to extract features in the frequency domain for every 100 ms. We only use the first 25 frequency components out of 128 FFT frequency bins as most of the energy of activities is in lower frequencies, and in this way, the feature vector does not become sparse.

First, we use random forest with 100 trees for classification of activities. To have a feature vector that contains enough information about an activity, the modified STFT bins are stacked together in a vector for every 2 s of activity; hence, every feature vector will be of length 1000. We also implemented other techniques such as SVM, logistic regression, and decision tree; however, random forest outperformed these techniques. The confusion matrix for random forest is shown in Table 1a and, as observed, decent performance can be obtained for some of the activities, but not for activities such as "Lie down," "Sit down," and "Stand up."

We also apply HMM on the extracted features using STFT and use the MATLAB toolbox for HMM training. Note that HMM is also used in CARM; however, DWT and the technique in [8] are used for feature extraction. The result is shown in Table 1b, and improved accuracy compared to random forest can be observed, although with higher computation time needed for training. Although the performance of HMM is good, especially for "Walk" and "Run," it sometimes misclassifies "Stand up" with "Sit down" or "Lie down."

We evaluate the performance of LSTM using Tensorflow in Python. The input feature vector is the raw CSI amplitude data, which is a 90-dimensional vector (3 antennas and 30 sub-carriers). The LSTM approach is different from conventional approaches in the sense that it does not use PCA and STFT, and can extract features from CSI directly. The number of hidden units is chosen to be 200 where we consider only one hidden layer. For numerical minimization of cross entropy, we use the stochastic gradient descent (SGD) with batch size of 200 and learning rate of 10^{-4}. Our result is shown in Table 1c, where the accuracy is over 75 percent for all activities. One of the drawbacks of using LSTM in this way is the long training time compared to HMM. However, using deep learning packages such as Tensorflow, one can also use GPUs and speed up the training. Once the LSTM is trained, the test can be done quickly.

Discussions

Effect of Environment Change on Performance: The CSI characteristics are not the same for different environments and different people. There are different techniques to reduce the influence of environments [9]. For instance, after using PCA, the first component mainly includes the CSI information due to stationary objects [9]. By discarding the first principal component, the information due to the mobile target is mainly captured. Therefore, using this technique, relatively similar features can be obtained for different environments. Other techniques such as STFT and DWT represent the speed of change in the multipaths, which is related to the speed of movement of human body parts. Although the same activities in different environments result in very different CSI characteristics, due to similarity in the change of signal reflections, similar features can be obtained for different environments and people using STFT or DWT [9].

(a) Random forest

		Predicted					
		Lie down	Fall	Walk	Run	Sit down	Stand up
Actual	Lie down	0.53	0.03	0.0	0.0	0.23	0.21
	Fall	0.15	0.60	0.03	0.07	0.1	0.05
	Walk	0.04	0.05	0.81	0.07	0.01	0.01
	Run	0.01	0.03	0.05	0.88	0.01	0.01
	Sit down	0.15	0.03	0.02	0.04	0.49	0.26
	Stand up	0.10	0.03	0.02	0.06	0.20	0.57

(b) Hidden Markov model

		Predicted					
		Lie down	Fall	Walk	Run	Sit down	Stand up
Actual	Lie down	0.52	0.08	0.08	0.16	0.12	0.04
	Fall	0.08	0.72	0.0	0.0	0.2	0.0
	Walk	0.0	0.04	0.92	0.04	0.0	0.0
	Run	0.0	0.0	0.04	0.96	0.0	0.0
	Sit down	0.0	0.04	0.0	0.0	0.76	0.20
	Stand up	0.16	0.04	0.0	0.0	0.28	0.52

(c) Long short-term memory

		Predicted					
		Lie down	Fall	Walk	Run	Sit down	Stand up
Actual	Lie down	0.95	0.01	0.01	0.01	0.00	0.02
	Fall	0.01	0.94	0.05	0.00	0.00	0.00
	Walk	0.00	0.01	0.93	0.04	0.01	0.01
	Run	0.00	0.00	0.02	0.97	0.01	0.00
	Sit down	0.03	0.01	0.05	0.02	0.81	0.07
	Stand up	0.01	0.00	0.03	0.05	0.07	0.83

Table 1. Confusion matrix.

Effect of Wi-Fi Transmission Rate on Performance: In order for the CSI to show noticeable changes due to the movement, the rate of transmission should be high enough (nearly 1 kHz) to capture activities that are done quickly. We have observed severely degraded performance of classification methods when sampling rate is around 50 Hz. Increasing the frame rate increases the number of samples, and hence the computational cost increases for de-noising and feature extraction. Increasing the frame rate may also not help further after some point because human movement speed is limited in indoor areas. Therefore, by selecting a suitable sampling rate (around

There are still several challenges that need to be addressed in future work such as how to use CSI phase information in addition to the amplitude, how to make the system robust in different dynamic environments, and how to identify the behaviors of multiple users.

1 kHz), a good trade-off between the computational cost and the accuracy can be obtained.

Using CSI Phase Information: Due to errors such as CFO and SFO, the phase of WiFi CSI has rarely been used for activity recognition in the literature. However, by subtracting the phase information of neighboring antennas from one another, the CFO and SFO are omitted. The phase difference is related to the angle of arrival (AOA), although there is integer ambiguity in the number of full cycles of the received signal. The change in the target location can change the AOA and hence the phase difference. When the movement is fast and drastic, the signal will be scattered by the human body more randomly, and hence the AOA and phase difference will change faster. It might thus be helpful to use phase difference together with amplitude for feature extraction and then apply classification algorithms. However, further investigation will be left for future work due to lack of space.

Multi-User Activity Recognition: While many activity recognition techniques have been tested for a single user, the more interesting and also challenging problem is the case where multiple people are in the environment. One solution has been proposed in [2] to use the idea of MIMO receivers to separate the signals due to two distinct mobile objects. Having multiple receivers might also help in distinguishing the activities of multiple users. Some techniques for multi-speaker recognition might be applicable to the activity recognition problem. This remains an interesting open problem.

Conclusion and Future Work

In this work, a survey of recent advancements in human activity recognition systems using WiFi channel has been provided. The literature in this area shows great promise in achieving good accuracy in indoor environments. Using numerical testing, it has been observed that better accuracy can be obtained by employing deep learning techniques such as RNN LSTM rather than methods such as HMM. There are still several challenges that need to be addressed in future work such as how to use CSI phase information in addition to the amplitude, how to make the system robust in different dynamic environments, and how to identify the behaviors of multiple users.

References

[1] C. Han *et al.*, "Wifall: Device-Free Fall Detection by Wireless Networks," *IEEE INFOCOM*, 2014, pp. 271–79.

[2] Q. Pu *et al.*, "Whole-Home Gesture Recognition Using Wireless Signals," *Proc. 19th ACM Annual Int'l. Conf. Mobile Computing and Networking*, 2013, pp. 27–38.

[3] G. Wang *et al.*, "We Can Hear You with Wi-Fi!" *IEEE Trans. Mobile Computing*, vol. 15, no. 11, Nov. 2016, pp. 2907–20.

[4] O. Politi, I. Mporas, and V. Megalooikonomou, "Human Motion Detection in Daily Activity Tasks Using Wearable Sensors," *Proc. 22nd IEEE Euro. Signal Processing Conf.*, 2014, pp. 2315–19.

[5] A. Tahat *et al.*, "A Look at the Recent Wireless Positioning Techniques with a Focus on Algorithms for Moving Receivers," *IEEE Access*, vol. 4, 2016, pp. 6652–80.

[6] J. Wilson and N. Patwari, "Radio Tomographic Imaging with Wireless Networks," *IEEE Trans. Mobile Computing*, vol. 9, no. 5, 2010, pp. 621–32.

[7] F. Adib *et al.*, "3D Tracking Via Body Radio Reflections," *11th USENIX Symp. Networked Systems Design and Implementation*, 2014, pp. 317–29.

[8] P. V. Dorp and F. Groen, "Feature-Based Human Motion Parameter Estimation with Radar," *IET Radar, Sonar & Navigation*, vol. 2, no. 2, 2008, pp. 135–45.

[9] W. Wang *et al.*, "Understanding and Modelling of WiFi Signal Based Human Activity Recognition," *Proc. 21st Annual Int'l. Conf. Mobile Computing and Net.*, 2015, pp. 65–76.

[10] X. Yang, "IEEE 802.11 n: Enhancements for Higher Throughput in Wireless LANs," *IEEE Wireless Commun.*, vol. 12, no. 6, 2005, pp. 82–91.

[11] D. Halperin *et al.*, "Tool Release: Gathering 802.11n Traces with Channel State Information," *ACM SIGCOMM CCR*, vol. 41, no. 1, Jan. 2011, p. 53.

[12] S. Sen *et al.*, "You Are Facing the Mona Lisa: Spot Localization Using PHY Layer Information," *Proc. 10th Int'l. Conf. Mobile Systems, Applications, and Services*, 2012, pp. 183–96.

[13] Y. Wang *et al.*, "E-Eyes: Device-Free Location-Oriented Activity identification Using Fine-Grained WiFi Signatures," *Proc. 20th Annual Int'l. Conf. Mobile Computing and Networking*, 2014, pp. 617–28.

[14] K. Ali *et al.*, "Keystroke Recognition Using WiFi Signals," *Proc. 21st Annual Int'l. Conf. Mobile Computing and Networking*, 2015, pp. 90–102.

[15] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," *IEEE Int'l. Conf. Acoustics, Speech and Signal Processing*, 2013, pp. 6645–49.

Biographies

Siamak Yousefi (siamak.yousefi@utoronto.ca) received his Ph.D. degree in electrical and computer engineering from McGill University in 2015. Since then he has been a postdoctoral fellow at the Department of Electrical and Computer Engineering, University of Toronto, Canada. He is a recipient of a postdoctoral grant from Fonds de recherche du Quebec-nature et technologies (FQRNT). His research interests include applications of statistical signal processing and machine learning techniques for indoor positioning and human activity recognition.

Hirokazu Narui (hirokaz2@stanford.edu) received his M.S. degree in physics and electronics from Osaka Prefecture University, Japan, in 2010. He is currently a visiting researcher in the Department of Computer Science at Stanford University while working at Furukawa Electric Co., Ltd. in Japan. His main research interests are machine learning, neural networks, and indoor positioning.

Sankalp Dayal (sankald@stanford.edu) received his B.S. in electrical engineering from the Indian Institute of Technology, Delhi, and his M.S. in electrical engineering from the University of California Santa Barbara, and is currently continuing his studies in artificial intelligence at Stanford University. His research interests include applied machine learning and signal processing for AI-based consumer applications. He holds a patent on human motion-based pattern matching. He is currently working as a senior algorithm engineer at STMicroelectronics.

Stefano Ermon (ermon@cs.stanford.edu) is an assistant professor in the Department of Computer Science at Stanford University, where he is affiliated with the Artificial Intelligence Laboratory and the Woods Institute for the Environment. He received his Ph.D. in computer science from Cornell University in 2015. He has co-authored over 40 publications, and has won several best paper awards at AAAI, UAI, and CP. He is a recipient of the NSF Career Award.

Shahrokh Valaee (valaee@ece.utoronto.ca) is a professor with the Department of Electrical and Computer Engineering, University of Toronto. From 2010 to 2012, he was an Associate Editor of *IEEE Signal Processing Letters*, and from 2010 to 2015 an Editor of *IEEE Transactions on Wireless Communications*. He is a Fellow of the Engineering Institute of Canada.

Wi-Fi Radar: Recognizing Human Behavior with Commodity Wi-Fi

Yongpan Zou, Weifeng Liu, Kaishun Wu, and Lionel M. Ni

A novel type of system based on commodity Wi-Fi is increasingly attracting intense academic interest. Without hardware modification and redeployment, researchers are exploiting channel state information output by commodity Wi-Fi and transforming existing Wi-Fi systems into radar-like ones that can recognize human behavior along with data communication.

ABSTRACT

Wi-Fi, which enables convenient wireless access to Internet services, has become integral to our modern lives. With widely-deployed Wi-Fi infrastructure, modern people can enjoy a variety of online services such as web browsing, online shopping, social interaction, and e-commerce almost at any time and any place. Traditionally, the most significant functionality of Wi-Fi is to enable high-throughput data communication between terminal devices and the Internet. However, beyond that, we observe that a novel type of system based on commodity Wi-Fi is increasingly attracting intense academic interest. Without hardware modification and redeployment, researchers are exploiting channel state information output by commodity Wi-Fi and transforming existing Wi-Fi systems into radar-like ones that can recognize human behavior along with data communication. This fancy functionality is tremendously expanding the boundaries of Wi-Fi to a new realm and triggering revolutionary applications in the context of the Internet of Things. In this article, we provide a guide to and introduce the impressive landscape of this new realm.

INTRODUCTION

Nowadays, Wi-Fi is widely used in our daily lives whether in private residential houses or public places such as libraries and offices. It provides users with convenient wireless access to online services such as information retrieval, social networking, and electronic commerce. Compared to other communication technologies, Wi-Fi holds the advantages of providing online services with higher data rates, greater mobility, and broader coverage. It is these positive properties that make Wi-Fi an attractive option with the explosive growth of mobile data traffic, and it is widely deployed in urban cities.

Technically, Wi-Fi refers to a wireless RF communication technology based on a family of standard protocols (e.g., IEEE 802.11 a/b/g/n/ac). It enables wireless data transfer between end devices and Internet infrastructure with high data rates. With the prosperity of the Internet of Things (IoT), connections between humans, objects, and the Internet are increasing more than ever before. A variety of IoT services have sprung up and brought about a sharp increase of data traffic. For example, location-based services (LBS), including mobile advertising, service recommendation, and weather alert, are appealing in most shopping malls, smart cities, and other scenarios. All of these indicate that Wi-Fi shall play a crucial role in the era of IoT.

However, this is far from the whole story of Wi-Fi and IoT. The advances in Wi-Fi physical (PHY) layer technologies have enabled users to get access to low-level information that portrays detailed characteristics of signal propagation through multipath, without hardware modifications or redeployment. A superior example of Wi-Fi PHY information is channel state information (CSI), which describes channel fading with amplitude and phase responses in fine granularity. The ever-increasing pervasiveness of Wi-Fi signals and growing fine-grained accessible PHY information not only open up new possibilities with Wi-Fi but also motivate researchers to pioneer a new realm beyond communication. In this article, we introduce such a pioneering area in which commodity Wi-Fi devices are transformed into radar-like systems that can sense and recognize human behavior by analyzing their motion states or gesture/activity types. *Informally, this kind of sensing system via Wi-Fi signals that achieves similar functions as radars is referred to as Wi-Fi radar in this article*. The significance of human behavior is two-fold. On one hand, it enables researchers to obtain better understanding of human behavioral patterns. On the other hand, recognizing human behavior has a wide range of applications in our lives, such as user authentication, healthcare for the elderly, location-based services, and human-device interaction, as shown in Fig. 1.

As is well known, radar is a kind of RF system that utilizes radio waves to sense, monitor, and track moving objects. With sophisticated hardware and high precision, radar is usually applied in certain specialized areas such as meteorology, the military, and traffic engineering. Similarly, *Wi-Fi radar* can also monitor targets' motions, and recognize human gestures and activities, but with lower precision and coarser granularity. Thus, a question naturally arises: *since radars have exceeding performance, what is the motivation behind designing Wi-Fi radar?* The reasons are mainly two-fold. For one thing, the requirements of sensing precision, granularity, and range for most IoT services are not as harsh as for radars. Take indoor location-based services (LBS) as an example. The acceptable precision of human motion monitoring is about meter level in typical indoor environ-

Yongpan Zou, Weifeng Liu, and Kaishun Wu are with Shenzhen University; Lionel M. Ni is with the University of Macau.

Digital Object Identifier:
10.1109/MCOM.2017.1700170

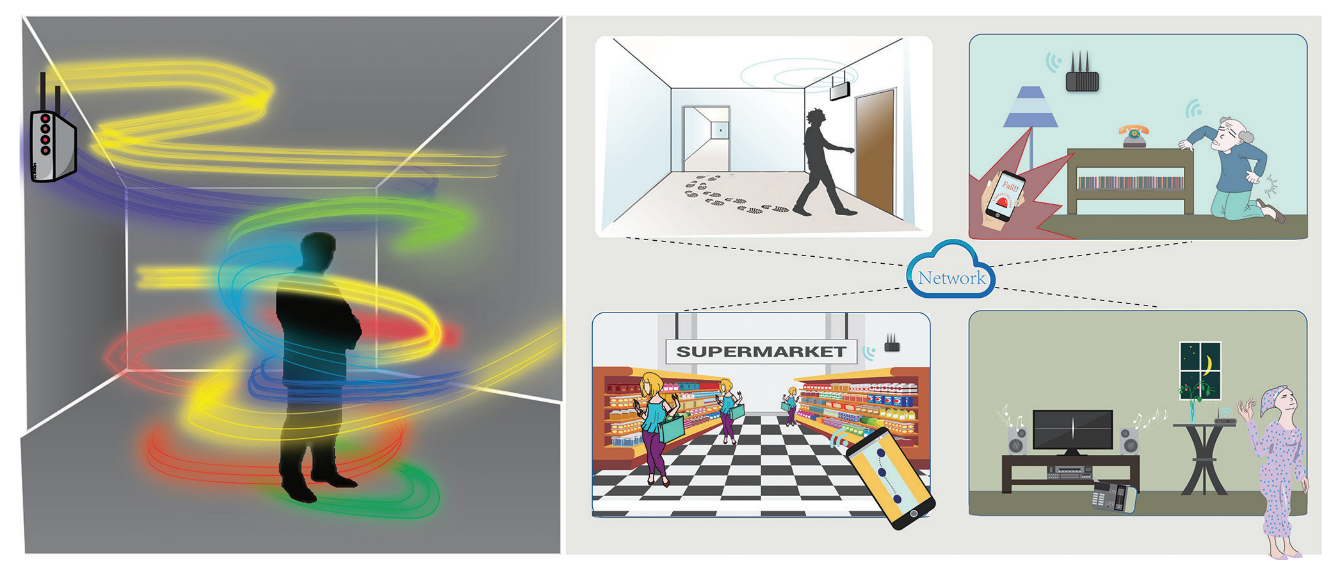


Figure 1. The visualization of Wi-Fi signals and application scenarios of Wi-Fi radar in IoT.

ments such as shopping malls, office spaces, and residential houses. For another thing, radar systems consist of costly and specialized hardware components, which makes them not as pervasive as Wi-Fi. The radar infrastructure and hardware components are rarely deployed in daily environments or embedded in commercial devices. In contrast, Wi-Fi infrastructure is widely deployed as aforementioned, and Wi-Fi Soc is pervasively embedded in various devices such as smartphones, tablets, smart TVs, and the like. Without hardware modification and redeployment, it is rather appealing to perform "secondary development" with commodity Wi-Fi and provide additional services along with communication.

Motivated by the inspiring prospect, researchers have made successive attempts to design Wi-Fi radar with CSI to recognize human behavior. In the remainder of this article, we first give an introduction to the knowledge background of Wi-Fi radar. Following that, an overview of the general design approaches and framework of Wi-Fi radar is presented. After this, a comprehensive survey of the state-of-the-art works in this field is conducted. At last, we put forward our remarks and conclusion.

Background Knowledge

As aforementioned, Wi-Fi radar is built on CSI. Before introducing Wi-Fi radar systems, it is better to give introduction to background knowledge closely related to CSI. In this section, we explain orthogonal frequency-division multiplexing (OFDM), based on which CSI is introduced in the following.

OFDM

OFDM represents a communication technology that transmits signals across orthogonal subcarriers at different frequencies. In OFDM, a wide frequency band is divided into multiple mutually orthogonal narrow subcarriers with different central frequencies. For data transmission, a high-rate bitstream in OFDM is first split into multiple relatively low-rate bitstreams, and then each of them is transmitted over a certain subcarrier independently. It is noted that, due to the orthogonality of subcarriers, bitstreams transmitted over multiple subcarriers simultaneously will not cause interference to each other. As a result, the spacing of subcarriers across the frequency band can be highly tight, which increases the spectrum efficiency to a great extent. In addition, multipath channel is usually frequency-selective, which means signals transmitted over subcarriers experience different levels of fading. Consequently, OFDM possesses better robustness to multipath interference (reflection, scattering, and absorption) than other transmission schemes, and thus is widely used in many wireless systems such as Wi-Fi and LTE.

Concept of CSI

In a Wi-Fi system with IEEE 802.11 a/g/n protocol, signals are transmitted in an OFDM scheme with $K = 48$ subcarriers. Due to the multipath effect, signals arriving at the receiver are not exactly the same as the original version from the transmitter. The differences are reflected in amplitude attenuation and phase shift. To accurately quantify such a channel fading effect, CSI is brought in originally for the sake of adapting transmission rate and optimizing throughput. Mathematically, for a system with M transmitting and N receiving antennas, each CSI sample at an instant is a collection of $M \times N$ complex vector $\mathbf{H} = [H(f_1), H(f_2), ..., H(f_k)]$, which describes the channel response at each subcarrier. Correspondingly, each element $H(f_k)$ in $\mathbf{H}$ represents the amplitude and phase response of the kth OFDM subcarrier that correlates the transmitted signal $X(f_k)$ and received signal $Y(f_k)$ in the frequency domain by $Y(f_k) = H(f_k)X(f_k)$. From the definition, it is obvious that due to OFDM, CSI can portray the propagation channel with a subcarrier-level granularity in the frequency domain and convey richer information compared to the summation of them (i.e., received signal strength indicator, RSSI). Figuratively speaking, CSI is to RSSI what a rainbow is to a sunbeam, where components of different frequencies are separated as shown in Fig. 2.

Properties of CSI

As aforementioned, CSI contains amplitude and phase response of the signal propagation channel at each subcarrier. In what follows, we shall drill down further and analyze the properties of CSI from the amplitude and phase perspectives. We expect that the analysis will shed light on principles of Wi-Fi radar.

Amplitude: By definition, the amplitude of CSI can easily be obtained by calculating the modulus of its every element (e.g., $||H(f_k)||$ for $H(f_k)$). The physical meaning of CSI amplitude is that it quantifies signal power attenuation after multipath fading. In this sense, it is similar to another signal indicator, RSSI. However, CSI amplitude has shown several favorable merits compared to other indicators:

- Frequency diversity. As mentioned above, CSI depicts channel response in each narrow subcarrier. Due to frequency-selective fading, signal streams on different subcarriers go through diverse multipath fading, which results in uncorrelated CSI values across subcarriers.
- Temporal stability. Since CSI amplitude is essentially a set of attenuation coefficients of a channel, it is rather robust to interference coming from transeceivers such as power adaptation, as long as there are no changes to the channel itself.
- Fine-grained granularity. Instead of measuring a channel with a composite value like RSSI, CSI decomposes a whole channel measurement into subcarriers and estimates the frequency response of each subcarrier, which obtains a finer-grained description of the channel in the frequency domain.

Phase: Phase is the information contained in CSI that is the counterpart to amplitude. Similarly, in order to extract phase information from CSI, we only need to calculate the angle of each complex element $\angle H(f_k)$. Theoretically, the phase of CSI has similar properties to amplitude. Nevertheless, the case is more complex for phase in practice. From another perspective, phase $\hat{\phi}_f$ extracted from CSI is composed of four different parts, that is, ϕ_f for genuine channel response phase, $2\pi f_f \Delta t$ for phase shift caused by clock offset, β for phase shift induced by carrier frequency offset, and Z_f for measurement noise. Since it is difficult to measure accurate clock and carrier frequency offsets on commodity devices, the phase extracted from raw CSI data is reported to be randomly distributed. Due to the randomness, the physical meaning of CSI phase is blurred, which means phase has rarely been utilized in previous work. Recently, some methods have been proposed to calibrate raw phase and treat calibrated phase as a new feature, but are still not accurate enough for modeling.

An Overview of a Wi-Fi Radar System

Wi-Fi radar has been informally defined in the introduction. However, the fundamental principles and general architecture of Wi-Fi radar remain unclear. In this section, we give more detail about the above aspects.

Human Behavior Recognition

Human behavior refers to the array of every physical action and observable emotion associated with individuals, and covers a wide range of specific contents. In the context of Wi-Fi sensing, present work toward behavior recognition can be classified mainly into three categories, namely, gesture recognition, activity monitoring, and motion tracking. Although it is difficult to give a precise definition to each category, there are notable differences among them from the perspectives of granularity and continuity. Intuitively, gestures only involve a certain part of the human body, such as finger, hand, arm, and even lip, and are of relatively short duration. In contrast, activities cover more body parts and consist of a sequence of physical actions. Human motion describes continuous physical movement of a whole body or just a certain part. And most of the time, motion tracking outputs human position and direction with high precision in real time.

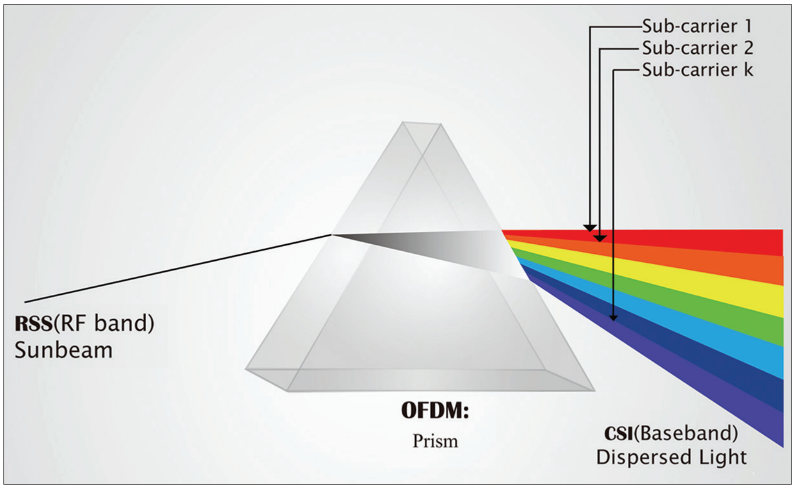


Figure 2. Illustration of the relationship between CSI and RSSI.

In fact, there are already some methods for behavior recognition, mainly including sensor-based and camera-based ones. Then what are the benefits of Wi-Fi radar that make it worthy of intense attention from researchers? Obviously, due to the pervasiveness of Wi-Fi infrastructure and devices, Wi-Fi radar has lower hardware cost than these two approaches since they both need additional devices. In addition, compared to a sensor-based approach, Wi-Fi radar works in a device-free way without requiring users to wear any sensors. This is more comfortable and convenient, especially in certain circumstances such as showering. In comparison with the camera-based approach, Wi-Fi radar is not dependent on line of sight (LOS) and light conditions.

Why Is CSI Feasible and Better?

As demonstrated above, CSI has several intrinsic advantages for communication purposes. But it is still unrevealed to readers why CSI is capable of capturing human behavior and better to utilize for designing Wi-Fi radar. As we all know, signals sent by a transmitter travel through multiple propagation paths and experience reflection and scattering before arriving at a receiver. Human behavior (e.g., gestures and activities) is bound to cause significant changes to the propagation channel of signals by altering the multipath. Since CSI quantifies the channel fading effect in amplitude attenuation and phase shift, it is sensitive to

Category	Layer	Resolution	Frequency diversity	Behavior sensitivity	Hardware accessibility
RSSI	MAC	Time domain: packet level Frequency domain: N/A	No	Low	Handy access
CSI	Physical	Time domain: multipath cluster Frequency domain: subcarrier level	Yes	High	Wi-Fi NIC

Table 1. The comparison between RSSI and CSI in human behavior recognition.

any changes of channel state, and thus is capable of capturing human behavior. Moreover, compared to other Wi-Fi indicators such as RSSI, a favorable point of CSI is that it depicts channel state with finer-grained frequency resolution and equivalently higher time resolution to distinguish multipath components as demonstrated by previous research [1]. In other words, it indicates that CSI possesses higher sensitivity to human behavior and is more powerful in uncovering behavior. The comparison between CSI and RSSI can be seen in Table 1.

General Approaches and Framework

In general, Wi-Fi radar systems in present works can be divided into two main streams according to their design approaches. One of them is the data-driven approach, which highly relies on collecting a large amount of data and adopts a training-learning scheme in a supervised or unsupervised way. The rationale lies in the fact that behavior causes changes to multipath and thus results in distinguishable patterns in CSI. By mining the patterns with machine learning techniques, it is possible to recognize behavior. However, without deterministic one-to-one mapping between CSI data and behavior, this approach can only recognize a set of predefined behavior in a certain system and is limited to application in gesture and activity recognition. The other one is the model-based approach. In this approach, deterministic models are built based on physical principles and correlate CSI with behavior with one-to-one mapping. Different from the data-driven approach, it can monitor human behavior continuously with little system training effort. However, due to the great challenge in modeling, there are only a few works that conduct such attempts. Based on the above, we can give a general framework of Wi-Fi radar, as shown in Fig. 3, which mainly consists of three layers: the hardware/infrastructure layer, the data processing layer, and the application layer. In the following, we give details about the above two main approaches.

How Is This Field Evolving?

To the best of our knowledge, [2] is the first work that makes use of CSI to replace RSSI for more accurate rate adaptation and higher throughput in data transmission. In this work, measurements have been conducted to verify the merits of CSI in temporal stability and frequency diversity. Inspired by the reported advantages, researchers have started to explore applying CSI in other areas beyond communication. Wu *et al.* [3] explicitly came up with CSI-based fine-grained indoor localization with commodity devices for the first time. The insights are two-fold, high spatial discrimination and resilience to transmission variation, brought about by CSI. Later, Han *et al.* [4] introduced CSI into the area of human behavior recognition by designing a fall detection system with CSI. The rationale of CSI-based behavior recognition is that CSI is a fine-grained feature sensitive to body movements. Within this area, Wang *et al.* [5] first proposed a model to estimate human walking speed and further utilize the model to recognize activities. Recently, Wang *et al.* [6] brought in a Fresnel model, a well-known physical model, to shed light on fundamental principles and push the limit of high-precision activity recognition with CSI.

A Survey of the State of the Art

As a novel and promising technology, Wi-Fi radar exhibits its attractiveness in various interesting applications. It is these applications that tremendously expand the boundaries of Wi-Fi functionalities and reshape our conventional knowledge of Wi-Fi. In order to demonstrate this clearly, we conduct a thorough survey of the state-of-the-art Wi-Fi radar systems and categorize them into two main streams according to their design approaches: data-driven and model-based.

The Data-Driven Approach

The underlying principle of the data-driven approach has been introduced earlier with a high-level idea about this approach. In this section, we shall introduce in detail the general architecture and state-of-the-art applications of Wi-Fi radar built on this approach.

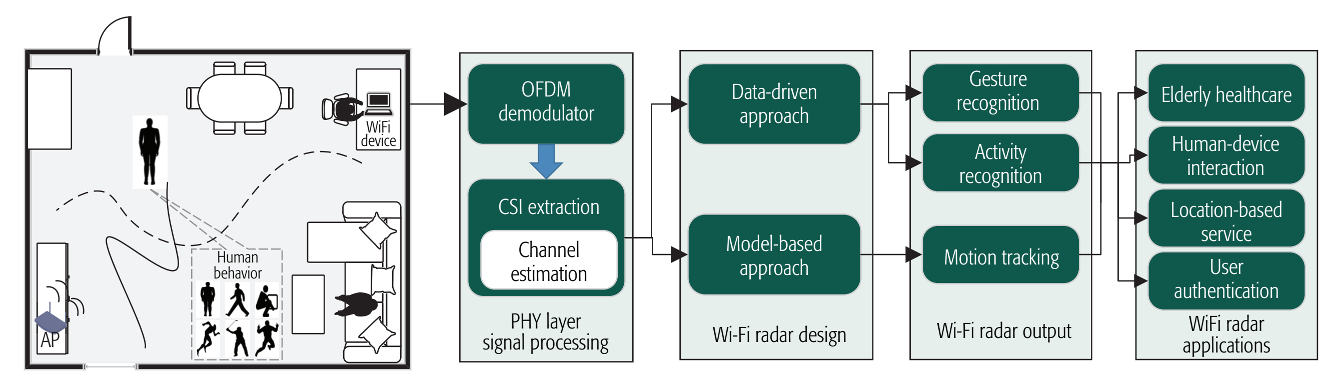


Figure 3. The general framework of a Wi-Fi radar system.

System Architecture: Even though the data-driven approach enables diverse Wi-Fi radar for specific purposes, a general architecture can be abstracted from these systems, as shown in Fig. 4. In general, the system architecture of a data-driven Wi-Fi radar consists of three stages, including data preprocessing, feature extraction and selection, and model training and testing, after CSI data is extracted from commodity Wi-Fi devices. The data flow in the architecture of Wi-Fi radar goes through nearly the same routine in other machine learning applications. However, considering the underlying physical meaning of CSI, there are some unique insights to be utilized in the system design, which are introduced in the following. To clean data in the first stage, researchers usually start by analyzing the properties of signals of interest in the time and frequency domains. Based on the results of signal analysis, they are able to design an appropriate denoising method to remove unwanted components as much as possible. Although the denoising techniques are customized case by case, there are some widely used methods such as Butterworth filtering and wavelet denoising due to their favorable properties [7]. Another important step in the preprocessing stage is behavior detection, which extracts signal segments corresponding to behavior events and discards the useless signals in a CSI sequence. This is essentially a signal detection problem in which energy-based hypothesis testing is frequently used to detect the start and end points of an event. Following data preprocessing, feature extraction and selection are performed on obtained segments. Although features are customized for different systems, they can roughly be classified into time domain, frequency domain, and time-frequency domain. Due to the lack of domain knowledge of the problem, it is common that researchers tend to extract features blindly in a preliminary attempt. Directly utilizing these features to train a machine learning model results in unsatisfactory performance and incurs heavy training overhead in most cases, since some features are noisy and redundant. As a result, a feature selection process is introduced to filter out those features and choose an optimal feature set that can achieve favorable performance and decrease training overhead. The common feature selection methods include correlation-based filtering, PCA, sequential search, and the like. In the last stage, selected features are fed into a certain machine learning model to construct predictive models in the training stage that are used for behavior recognition in the later learning process. The support vector machine (SVM), random forests, and hidden Markov model (HMM) are the common machine learning models in present data-driven Wi-Fi radar.

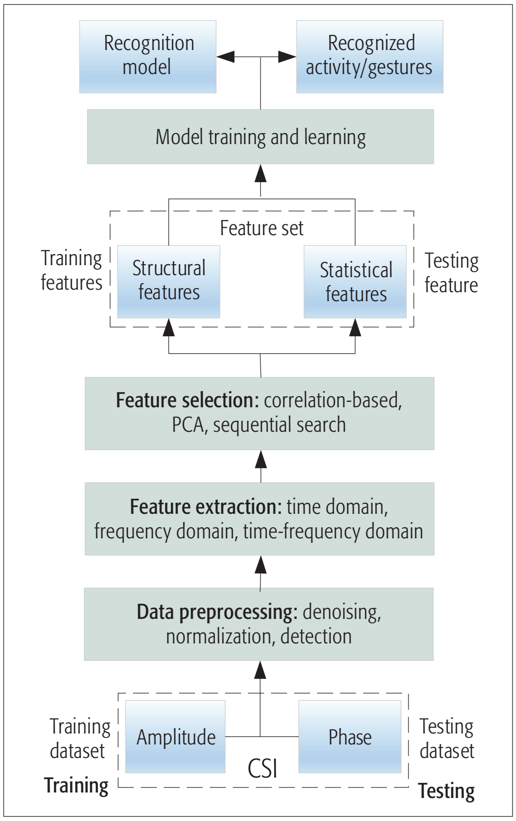

Recognition model
Recognized activity/gestures
Model training and learning
Feature set
Training features
Structural features
Statistical features
Testing feature
Feature selection: correlation-based, PCA, sequential search
Feature extraction: time domain, frequency domain, time-frequency domain
Data preprocessing: denoising, normalization, detection
Training dataset
Amplitude
Phase
CSI
Testing dataset
Training
Testing

Figure 4. The general architecture of data-driven Wi-Fi radar.

Another important step in the preprocessing stage is behavior detection, which extracts signal segments corresponding to behavior events and discards the rest useless signals in a CSI sequence. This is essentially a signal detection problem in which energy-based hypothesis testing is frequently used to detect the start and end points of an event.

Applications: As aforementioned, data-driven Wi-Fi radar is mainly applied for human gesture and activity recognition. Within this scope, a number of systems have been developed to recognize a variety of gestures or activities with various granularities. According to our survey, WiFall [4] and E-eyes [8] are the first works that bring CSI in the area of human activities recognition. Nevertheless, they have different focuses. WiFall concentrates on detecting whole-body activities such as falling, walking, and sitting of the elderly in case of emergency. By identifying the unique changes of CSI, the system is able to differentiate three activities with high accuracy. On the other hand, E-eyes focuses on recognizing human daily activities in residential houses such as cooking, brushing, bathing, and watching TV. Compared to WiFall, activities in E-eyes possess higher diversity and complexity. Later works within this scope mainly attempt to design Wi-Fi radar systems to recognize finer-grained activities such as smoking with more sophisticated system design [9]. Another application area of data-driven Wi-Fi radar is gesture recognition. Wang *et al.* [7] first proposed Wihear to recognize minute lip gestures with CSI when someone is speaking. By mapping lip gestures with vowels and consonants, they finally recover what the person has said. However, to prevent external interference from overwhelming signals of interest, Wihear adopts directional antennas and only works in a highly controlled environment. Later research in CSI-based gesture recognition moved toward recognizing gestures with commodity devices in a more realistic environment, covering a variety of applications such as human-computer interaction (HCI) [10], vital signs monitoring [11], keystroke eavesdropping [12], and user authentication [13].

The Model-Based Approach

In addition to the data-driven approach, Wi-Fi radar can also be designed with deterministic physical models and track human motion in real time. Although such Wi-Fi radar systems possess advantages of real-time monitoring and little system training, the critical challenge lies in developing suitable models that correlate CSI with human behavior in a one-to-one mapping. Due to this

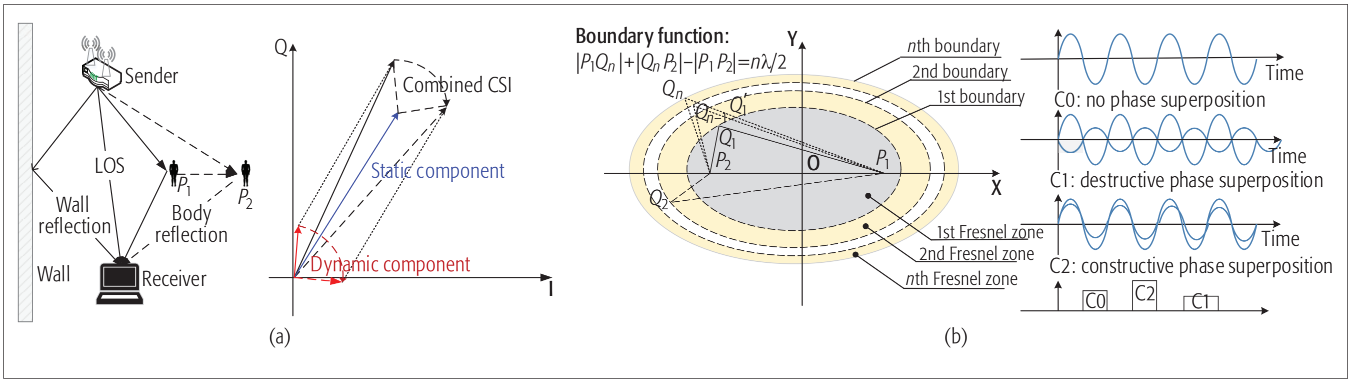


Figure 5. Illustration of a CSI-speed model and a Fresnel zones model: a) visual and phasor representations of the CSI-speed model; b) the Fresnel zone model and corresponding signal superposition.

challenge, only a few works have been conducted in this direction. In the following, we mainly introduce two different models developed to monitor human walking and/or vital signs.

CSI-Speed Model: In CARM [5], the authors propose a CSI-based speed model that estimates the speed of human activities by monitoring the amplitude of CSI. The underlying principle of this model can be demonstrated as follows. In the setting shown on the left side of Fig. 5a, the receiver receives multipath components of transmitted signals, including the LOS component, wall and body reflections, and traveling from different propagation paths. In a static environment, where all the objects are static except the human body, the LOS component and wall reflection are static, and thus result in a constant I-Q vector in the complex plane, as shown in the right part of Fig. 5a. When a person moves from P_1 to P_2, the traveling distance of body reflection changes approximately by $|P_1P_2|$, which consequently induces a phase shift to body reflection. As a result, the composite phase of received signals is to be changed accordingly. In other words, the phase of received signals varies along with walking distance $|P_1P_2|$. Due to the randomness in CSI phase, the authors turn to tracking the variance of CSI amplitude instead of phase in the model, and build up a relationship between CSI amplitude and walking speed.

Fresnel Zones Model: Fresnel zones, named for physicist Augustin-Jean Fresnel, refer to a series of concentric prolate ellipsoidal regions between and around a pair of signal (acoustic or RF) transmitter and receiver (at P_1 and P_2 in Fig. 5b). This model is originally put forward to understand and compute the strength of waves propagating in the space. Referring to Fig. 5b, the innermost ellipse is defined as the first Fresnel zone, the elliptical annuli between the first ellipse and the second one is defined as the second Fresnel zone, and the Nth Fresnel zone corresponds to the elliptical annuli between the $(n - 1)$th and nth ellipses. Correspondingly, the boundary between any two adjacent Fresnel zones (say nth and $(n + 1)$th) is defined as the nth Fresnel boundary. Now assume the transmitter at P_1 is transmitting signals, and the receiver at P_2 only receives the LOS component through P_1P$_2$, if there are no other objects in the space (free space case). However, when a reflector is located in a position, say Q_1, the received signals are the combination of LOS component via P_1P_2 and multipath component via P_1Q_1 and Q_1P_2. Since these two components possess phase shifts of $2\pi f|P_1P_2|/c$, and $2\pi f(|P_1Q_1| + |Q_1P_2|)/c + \pi$ (c is signal speed, the added π phase shift is caused by reflection), the combination of them depends on their phase difference $\Delta\phi$ (i.e., $2\pi f/c(|P_1Q_1| + |Q_1P_2| - |P_1P_2|) + \pi$). When phase difference $\Delta\phi$ equals $k\pi$ and k is an odd number, indicating that phases of the two components are inverse, both components cancel each other, and the strength of received signals (i.e., CSI amplitude) decreases consequently. On the contrary, if k is an even number, both components reinforce each other, and thus strengthen the strength of received signals. In both cases, the reflector is certain to be located in the Fresnel boundary, which is mathematically determined by the equation shown in Fig. 5b. Moreover, when the reflector is located in a Fresnel zone, the weaken or strengthen effect is in between compared to the case where the reflector is on the corresponding boundary.

According to the above analysis, it is clear that the strength of received signals is highly sensitive to the location of the reflector. In other words, if we treat the human body as a reflector, since any body motion results in variance of CSI amplitude, it is feasible to detect the position and monitor the movement of a human body in real time by analyzing CSI data. Moreover, as the resolution of the Fresnel zone for Wi-Fi signals is centimeter-level, it is possible to achieve fine-grained body motion and activity tracking with this model [14]. Based on this insight, Wang *et al.* first made use of the Fresnel zones model to explain the fundamental principles of CSI-based human activity recognition, and further explored the effect of position and orientation on the performance of vital signs detection [6]. With this model, they also achieved high-precision walking direction estimation in [15].

Remaining Challenges and Open Issues

Wi-Fi radar is a promising technology that stands out for low hardware cost, pervasiveness, and unobtrusiveness. Although researchers have conducted pioneering exploration with notable achievements, there remain several critical challenges and open issues to be handled in order to further advance this area. First, for data-driven Wi-Fi radar, improving the system scalability is a big challenge. Existing systems following this design routine are usually required to be trained and tested in the same environment. It is questionable whether they can maintain high performance when the testing environment is changed even by furniture repositioning or other objects' presence. Second, for existing model-based systems, robustness is a major concern since multipath has a great effect on model performance. When motions or objects out of interest exist, how to remove the induced multipath interference and maintain performance is still a remaining challenge. Moreover, we hold the viewpoint that it is worth trying a combination of both approaches. More specifically, it is desirable to combine the physical meaning of CSI with data mining techniques in designing Wi-Fi radar systems.

Conclusion

In this article, we introduce an emerging and promising technology that transforms commodity Wi-Fi into radar-like systems that can recognize human behavior with channel state information. By surveying the latest works, we summarize the general framework of existing Wi-Fi radar systems, and figure out that the design of these systems mainly follows a data-driven approach or a model-based approach. For each kind of Wi-Fi radar, we give a detailed introduction to the fundamental principles and state-of-the-art applications. We envision that, although Wi-Fi radar still faces critical challenges toward practical use, it has shown a great vision of the future of the Internet of Things.

Acknowledgment

This research was supported in part by China NSFC Grant 61472259, Guangdong Natural Science Foundation No. 2017A030312008, Shenzhen Science and Technology Foundation (No. JCYJ20170302140946299, JCYJ20170412110753954). Guangdong Talent Project 2014TQ01X238, 2015TX01X111, GDUPS(2015), Hong Kong RGC Grant HKUST16207714, Macao FDCT Grant 149/2016/A and the University of Macau Grant SRG2015-00050-FST. Kaishun Wu is the corresponding author.

References

[1] Z. Yang, Z. Zhou, and Y. Liu, "From RSSI to CSI: Indoor Localization Via Channel Response," *ACM Computing Surveys*, vol. 46, no. 2, 2013, p. 25.

[2] D. Halperin *et al.*, "Predictable 802.11 Packet Delivery From Wireless Channel Measurements," *ACM SIGCOMM Computer Commun. Rev.*, vol. 40, no. 4, 2010, pp. 159–70.

[3] K. Wu *et al.*, "Fila: Fine-Grained Indoor Localization," *Proc. IEEE INFOCOM*, 2012, pp. 2210–18.

[4] Y. Wang, K. Wu, and L. M. Ni, "WiFall: Device-Free Fall Detection by Wireless Networks," *IEEE Trans. Mobile Computing*, vol. 16, no. 2, 2017, pp. 581–94.

[5] W. Wang *et al.*, "Understanding and Modeling of WiFi Signal Based Human Activity Recognition," *Proc. ACM Mobicom*, 2015, pp. 65–76.

[6] H. Wang *et al.*, "Human Respiration Detection with Commodity WiFi Devices: Do User Location and Body Orientation Matter?" *Proc. ACM Ubicomp*, 2016, pp. 25–36.

[7] G. Wang *et al.*, "We Can Hear You with Wi-Fi!," *IEEE Trans. Mobile Computing*, vol. 15, no. 11, 2016, pp. 2907–20.

[8] Y. Wang *et al.*, "E-Eyes: Device-Free Location-Oriented Activity Identification Using Fine-Grained WiFi Signatures," *Proc. ACM Mobicom*, 2014, pp. 617–28.

[9] X. Zheng *et al.*, "Smokey: Ubiquitous Smoking Detection with Commercial WiFi Infrastructures," *Proc. IEEE INFOCOM*, 2016, pp. 1–9.

[10] H. Li *et al.*, "WiFinger: Talk to Your Smart Devices with Finger-Grained Gesture," *Proc. ACM Ubicomp*, 2016, pp. 250–61.

[11] X. Liu *et al.*, "Wi-Sleep: Contactless Sleep Monitoring via WiFi Signals," *Proc. IEEE RTSS*, 2014, pp. 346–55.

[12] K. Ali *et al.*, "Keystroke Recognition Using WiFi Signals," *Proc. ACM Mobicom*, 2015, pp. 90–102.

[13] T. Xin *et al.*, "Freesense: Indoor Human Identification with Wi-Fi Signals," *Proc. IEEE GLOBECOM*, 2016, pp. 1–7.

[14] D. Zhang, H. Wang, and D. Wu, "Toward Centimeterscale Human Activity Sensing with Wi-Fi Signals," *IEEE Computer*, vol. 50, no. 1, 2017, pp. 48–57.

[15] D. Wu *et al.*, "Widir: Walking Direction Estimation Using Wireless Signals," *Proc. ACM Ubicomp*, 2016, pp. 351–62.

Biographies

Yongpan Zou (yongpanzou2012@gmail.com) is now an assistant professor at Shenzhen University (SZU). He obtained his Ph.D. degree in 2017 from the Department of Computer Science and Engineering of Hong Kong University of Science and Technology (HKUST). In 2013, he received his B.Eng. degree in chemical machinery from Xian Jiaotong University, Xian, China. He received the Best Paper Award of MASS 2014. His current research interests mainly include mobile computing, embedded systems, and wireless sensing.

Weifeng Liu (szuliuweifeng@gmail.com) received his B.Eng. degree in computer science from Shenyang Ligong University, China. Since 2015 he has been a Master's student in the College of Computer Science and Software Engineering at SZU. His current research interests mainly include mobile computing, wireless sensing, and data mining.

Kaishun Wu (wu@szu.edu.cn) received his Ph.D. degree in computer science and engineering from HKUST in 2011. After that, he worked as a research assistant professor at HKUST. In 2013, he joined SZU as a professor. He is the inventor of 6 US and 43 Chinese pending patents (13 are issued). He received the 2014 IEEE ComSoc Asia-Pacic Outstanding Young Researcher Award.

Lionel M. Ni (ni@umac.mo) is Chair Professor in the Department of Computer and Information Science and Vice Rector of Academic Affairs at the University of Macau. Previously, he was Chair Professor of Computer Science and Engineering at HKUST. He received his Ph.D. degree in electrical and computer engineering from Purdue University in 1980. He has chaired over 30 professional conferences and received eight awards for authoring outstanding papers.

Wi-Fi radar is a promising technology that stands out by low hardware cost, pervasiveness and unobtrusiveness. Although researchers have conducted pioneering exploration with notable achievements, there still remain several critical challenges and open issues to be handled in order to further advance this area.

BEHAVIOR RECOGNITION BASED ON WI-FI CSI

Human Behavior Recognition Using Wi-Fi CSI: Challenges and Opportunities

Lili Chen, Xiaojiang Chen, Ligang Ni, Yao Peng, and Dingyi Fang

The authors present a comprehensive introduction to HBR using Wi-Fi channel state information. They review the state-of-art of HBR, based on the two techniques that drove recent progress in Wi-Fi channel-state-information-based HBR: fingerprint-based and model-based. Specifically, they describe their corresponding characteristics, general architectures, and provide a performance comparison of the two mechanisms.

ABSTRACT

Human behavior recognition (HBR) has emerged as a core research area in human-computer interaction. In this article, we give a comprehensive introduction to HBR using Wi-Fi channel state information. We first comprehensively review the state-of-art of HBR, based on the two techniques that drove recent progress in Wi-Fi channel-state-information-based HBR -- fingerprint-based and model-based. Specifically, we describe their corresponding characteristics, general architectures, and provide a performance comparison of the two mechanisms. We then provide insights into the future directions of HBR research, and propose two possible new schemes, and the technical challenges coming with them.

INTRODUCTION

Device-free human behavior recognition (HBR), that is, automatically recognizing physical behaviors without attaching sensors [1] or tags [2] on human bodies, acts as a key enabler for various applications including augmented reality games, smart homes, remote healthcare, and so on. Nowadays, although the device-free HBR techniques have considerably matured, there are still a number of challenges and opportunities.

As illustrated in Table 1, traditional approaches using visual [3, 4] and acoustic [5, 6] signals have had commercial success. However, computer-vision-based techniques have the fundamental limitations of requiring a line-of-sight (LOS) path to operate and potentially leaking human privacy. Sound-based methods aim for minus behavior recognition due to the short wavelength of sound signal, thus limiting their usage within a small area.

To overcome the above limitations, the recent ubiquity of Wi-Fi-enabled devices has inspired many Wi-Fi-based HBR interfaces [7–15] that cover multi-granularity behavior recognition and can operate in non-LOS (NLOS) scenarios. The intuition is that a certain human behavior introduces unique multi-path reflection in Wi-Fi signals, as shown in Fig. 1. Through a comprehensive survey on recent advances, we find that the state of the art on Wi-Fi-based HBR mainly involves the following two schemes.

The first uses a single wireless monitor to detect and identify human behavior [7–9], such as modest WiVi [7] or fine-grained WiSee [8] and WiTrack [9]. The key advantage of these systems is that they require no training for recognition, as they establish the signal-behavior models by making full use of the phase and amplitude dynamics in wireless signals. Unfortunately, these systems are all equipped with a specialized Wi-Fi monitor (e.g., USRP) for extracting the carrier wave features that are not reported in current Wi-Fi systems. Overcoming this limitation, another kind of HBR system has been developed built with inexpensive commercial off-the-shelf (COTS) Wi-Fi devices [10-15]. These systems analyze human behavior's effect on Wi-Fi signals. There are two kinds of commonly used Wi-Fi signal information: channel state information (CSI) [10–14] and received signal strength indicator (RSSI) [15]. CSI has received more attention than RSSI in the HBR research community because CSI provides information that is more suitable for HBR. CSI contains amplitude and phase information needed by HBR, while RSSI has only average signal strength value for each packet.

In this article, we review the recent progress in Wi-Fi CSI-based HBR and propose promising directions. The prior works can be classified into two categories: model-based [13, 14] and fingerprint-based [10–12], depending on whether they require a priori learning. Model-based systems recognize behaviors without training, but require a large number of access points (APs) to accurately track human body trajectory, making them impractical in many applications. Alternatively, fingerprint-based methods can work with just a single AP, but at the cost of complex and often time-consuming pattern matching, prohibiting real-time operations.

Having surveyed the state of the art of HBR techniques, we first propose several potentially promising research directions. We then suggest innovative solutions for HBR, including deep learning and signal modeling, and discuss a number of challenges: eliminating the multi-path effect in real environments, dealing with the time asynchronous problem in COTS Wi-Fi devices, extracting tiny reflected signals from mixed measurements in receivers, improving the validity of Fresnel zone boundaries detection, and distinguishing simultaneous behaviors of multiple body parts.

Contributions: We are the first to provide a tutorial on Wi-Fi CSI-based HBR. The strengths lie in proposing constructive insights for future trends and research opportunities, and concluding with prior works comprehensively. This will inspire newcomers to quickly understand the state of the art and develop more practical interfaces while building on poor infrastructure.

Digital Object Identifier:
10.1109/MCOM.2017.1700081

The authors are with Northwest University.

Wi-Fi CSI-Based Behavior Recognition Approaches

Recently, CSI is gaining more popularity in various multi-granularity behavior recognition applications, including lip language, drawings, body motion, falls, and so on [10–14]. Despite the diverse granularity in these systems, we can classify these interfaces into two classes: model-based and fingerprint-based behavior recognition, depending on whether they require a priori learning. In the following subsections, we provide a detailed description of the two classes and compare their performance.

Understanding of Two Categories of Schemes

Most Wi-Fi CSI-based HBR systems achieve behavior recognition by fingerprint matching. The intuition is that different human behaviors cause the received CSI streams to generate different dynamics, which can be utilized to construct profiles for predefined behaviors in an offline training phase. In an online recognition phase, the interface will search for a best-fit profile for the performed behavior through minutiae matching. The key benefit of fingerprint-based recognition systems is that they do not require intensive AP deployment, and can work with even a single AP. This ensures low hardware cost and maintenance, as well as almost no inconvenience in a person's normal life. Therefore, fingerprint-based schemes are more suitable for scenarios where there are densely populated and frequent activities.

Unlike fingerprint-based schemes that all require a priori learning of CSI measurement patterns, which limit them to recognizing only a fixed set of pre-defined human behaviors, model-based schemes do not need to pre-define behaviors. Model-based schemes utilize two basic ideas:

1. Whenever the user's moving body blocks a signal coming from a certain direction, the signal strength of the angle of arrival (AoA) representing the same direction experiences a sharp drop. Thus, the interface can track the user's body parts trajectory and further identify the user's behavior by leveraging the AoA values of incoming CSI streams at the mobile device.
2. The Fresnel model for WiFi radio propagation also works in a multi-path environment. By exploiting the Fresnel model and WiFi radio propagation properties deduced, it is feasible to investigate the impact of human behavior on the receiving CSI streams, and develop a theory to correlate one's body part's moving distance, location, and orientation to the identifiability of human behavior.

Since the model-based interfaces require no a priori learning, they can track an arbitrary set of body motions, which enables wider ranges of real-time applications, especially for virtual handwriting application.

Overview Architecture Of Wi-Fi CSI Based HBR

The overall architecture of Wi-Fi CSI-based HBR involves signal processing, model calculation, and machine learning techniques. In this section, we describe the general architecture of interfaces with two categories:- fingerprint-based and model-based. As shown in Fig. 2, it is obvious that the common component of the two types is their inputs, which all consist of CSI streams collected by COTS Wi-Fi devices. We introduce the specific components of each class in the following subsections.

Device-free HBR techniques				Features		
Vision-based				Works in LOS, easy to leak user privacy		
Sound-based				For minus behavior		
Wireless-signal-based	Specialized device			No training, specialized		NLOS, for multi-grained behavior
	COTS device	RSS		With training, coarse, only provides an average signal strength value		
		CSI	Fingerprint-based	With training, Sparse deployment	Fine-grained, provides detailed characteristics of the wireless link in the frequency domain, more popular than RSS	
			Model-based	No training, Intensive deployment		

Table 1. Comparison of device-free HBR techniques.

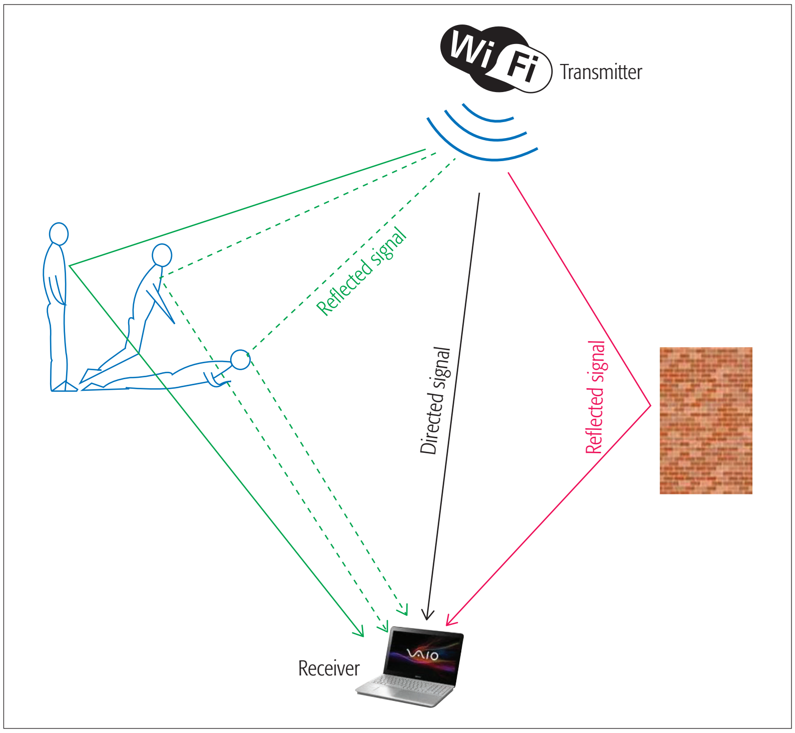


Figure 1. Multi-path environment with human behavior.

Fingerprint-Based:

CSI Preprocessing: Raw CSI measurements from commodity Wi-Fi devices are noisy due to multi-path propagation and hardware noise, which may cause false edges that affect the accuracy of HBR. Typical signal preprocessing mechanisms such as low-pass filters, principal component analysis (PCA), and discrete wavelet transform (DWT) are frequently applied to address this problem. By the principle that there is a difference between CSI dynamics caused by noise and human behavior, they can filter out the interference noise while preserving useful behavior information details in a CSI stream. This is essential for acquiring correct behavior segments.

Behavior Segmentation: The premise of cor-

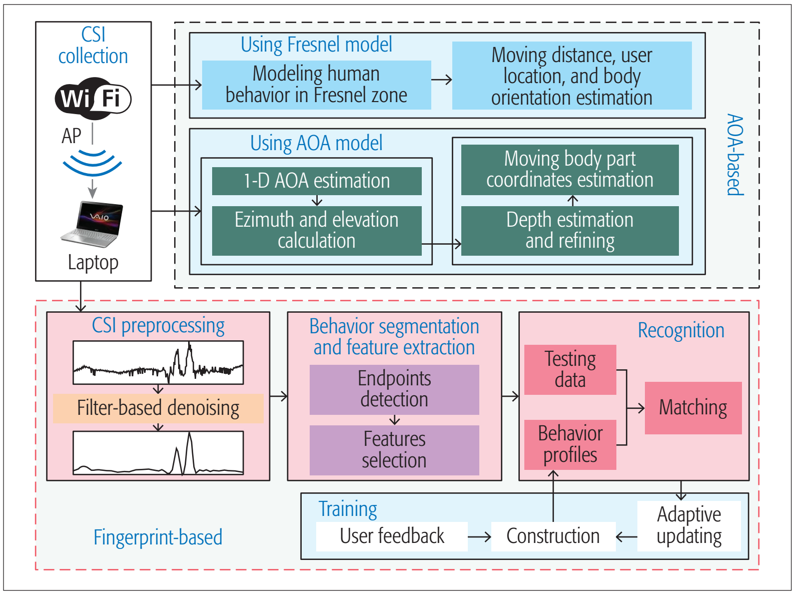


Figure 2. Overall architecture of Wi-Fi CSI-based HBR.

rect behavior recognition is to know when the user is performing a behavior. To answer this question, the behavior segmentation stage is involved to identify those segments of the preprocessed CSI streams that are likely to contain signs on behalf of behaviors. The intuition is that the effects of human behaviors on the received signal are either rising edges, falling edges, or pause. General behavior segmentation techniques are based on some special observed preambles that are hard to confuse with other actions in the environment. The preambles always contain two or three states such as drop and up-down in the input smooth CSI streams, which corresponds to a behavior period from beginning to end. For some slight behavior signals, successful behavior detection is due to an amplification technique at first. Then the behavior endpoints are detected by an effective threshold that is always updated adaptively. Finally, some special restrictions are added to exclude the false segments according to the respective characteristics of different behaviors.

Feature Extraction: Since the behavior segment always consists of a variety of data, it will be time-consuming if the segment is directly used for pattern matching. To address this problem, the feature extraction stage is significant, reducing the behavior segments into features that are discriminative for the behaviors. During this stage, features may be calculated automatically or derived based on expert knowledge. Past works specified the feature extraction stage as extracting the frequency components or calculating various metrics of amplitude from different behaviors at different timescales. Ideally, features corresponding to the same behavior are clustered in feature space, while features corresponding to different behaviors should be far apart. In addition, highly representative features need to be robust to different users as well as to intra-group variability of a behavior. Particularly for real-time processing on embedded systems, it is essential to make a reasonable trade-off between feature number and recognition accuracy. This is because of higher dimensionality of the feature space and higher classification accuracy, but greater computational complexity.

Training and Recognition: After extracting features from the behavior segment, the last significant step is to identify the behavior by a recognition algorithm, which includes two phases: training and recognition. In the training phase, the interface aims to establish behavior profiles using a large number of data that contains the characteristics of different behaviors. In the recognition phase, the best-fit profile is searched for a performed behavior by matching between the behavior segment features and the pre-constructed behavior profiles.

Research in computational statistics and machine learning developed a large number of recognition algorithms, which have been successfully applied to HBR. Typical template-based matching methods such as Dynamic Time Warping (DTW) can effectively classify the same behavior with various speeds:- one can wave slowly or quickly, but the behavior is just wave. For more complex data exhibiting temporal dependencies, temporal probabilistic models such as hidden Markov models (HMMs) and conditional restricted Boltzmann machines (CRBMs) have been widely used. In addition, discriminative approaches such as the support vector machine (SVM) also enable various HBR systems.

Model-Based: Model-based HBR using Wi-Fi CSI relies on two models: the Fresnel model and AOA model.

Using the Fresnel Model: The system using the Fresnel model consists of two stages. The aim of the first stage is to construct the Fresnel model of human behavior. The system usually models a human as a varying-size polyhedron simulating the body movements during behavior. Then, in the second stage, the system analyzes how the moving depth, user location, and body orientation of a human body affect the receiving CSI, and outputs the behavior that human has performed. The intuition is that a moving object usually creates a reflected signal with varying phase and amplitude. Within a small moving scope, the reflected signal generally has fixed amplitude and changing phase affecting the received signal. Within a large moving scale, the reflected signal suffers both amplitude variation and phase change as components of the received signal. Specifically, the system first converts the body displacement to the change of the reflected path length. Then the system converts the path length change to a CSI phase change. Finally, the CSI amplitude change is mapped uniquely to user location and body orientation. One point worth noting is that within each Fresnel zone, the worst human location for behavior perception is near the boundary, while the best location appears in the middle of the zone.

Using the AOA Model: The interface using the AOA model also involves two stages. During the first stage, the receiver leverages the CSI of incoming signals to compute their 1D AOA values by spatial spectrum estimation. The basic idea is that whenever the user's body part blocks the CSI measurement coming from a certain direction, the signal strength of the AOA corresponding to the same direction will drop significantly. Subsequently, the interface calculates the azimuth and elevation values from estimated 1D AOA output. During the second stage, the receiver first periodically collects RSSI and CSI from incoming signals. Then it uses CSI correlation values over time and filtering to filter out signals from mobile clients or those affected by large environmental variations. For stable AOAs, the interface processes the RSSI value and obtains a refined estimate of the moving distance. Finally, if the azimuth and elevation of the AOA are both known, the interface uses an algorithm to estimate the moving object's coordinates. For unknown AOAs, the moving object's coordinates are estimated as the weighted average of the coordinates from two known neighboring AOAs.

Performance Comparison of Selected Schemes

To evaluate different Wi-Fi CSI-based HBR schemes, Table 2 provides a detailed performance comparison for a selected set of interfaces [10–14]. For each interface, we choose the research that we perceived as the best, considering characteristics that are commonly provided in the literature such as behavior scale, average recognition accuracy, devices equipped, environment, and participants.

A major difficulty we faced during this comparison was the lack of standard evaluation parameters and environments used within the recognition research community. Most experiments were conducted in custom, controlled environments. From Table 2, we observe that:

1. The compared schemes are aimed at different types of behaviors' recognition, such as in-place, walking, or running activities, handwriting, respiration, and lip language. Different characteristics may affect the accuracy of recognition.
2. The environment in which the experiments were carried out has a major impact on the accuracy due to the distinctive multi-path distribution.
3. In addition, the compared schemes are across different fields, and the accuracy is affected by the number of devices equipped in the scenarios. Hence, the performances are almost incomparable without standard metrics or settings.

However, the comparison provided here can still serve as a good reference.

Future Trends and Research Opportunities

In view of the current research progress and practical application, we expect the future research in HBR to follow two trends.

Tap into Diverse Signals

The research on human behavior perception has made remarkable achievements by using acoustic, optical, RF, and WiFi signals. Thus, future research may tap more potential signals for behavioral recognition. Beyond amplitude and phase information in WiFi CSI, other information such as frequency and Doppler shift will also be explored to achieve behavior recognition by modifying 802.11 standards or using an extra monitor within the bounds of certain costs. The richness of information will make HBR more able to match the requirements of low cost, desired accuracy, high robustness, and so on. In a word, the signal diversity will create great opportunities for Wi-Fi CSI-based HBR.

Selected features	Fingerprint-based			Model-based	
	CARM	WiFall	WiHear	Respiration	WiDraw
Average accuracy	96%	87%	91%	cm level	91%
Participants	25	1	3	2	3
Devices	1	2	1	1	25
Scenario	Laboratory	Dormitory	Lobby	Office room	Office building
Behavior scale	In place, and walking and running	In place and walking	Lip language	Respiration	Handwriting

Table 2. Comparison of selected features from the best available recognition interfaces.

Multi-Grain-Oriented Behavior Recognition

To better understand multi-grain-oriented behavior recognition, we introduce the similar and maturely developed problem: gesture recognition. However, the interface for behavior identification differs from gesture recognition in that the interface needs to identify a series of multi-grained movements over a period of time rather than a single-grained and instantaneous body movement. For instance, a behavior such as talking includes fine-grain speaking and smiling, medium-grain head movements, coarse-grain walking, and so on. In conclusion, a behavior always covers multiple continuous gestures with different granularity. Therefore, it is imperative to expand the recognition technique for precise single gestures to multi-grain-oriented behavior recognition approach.

In order to realize the above mentioned recognition, deep learning and signal modeling will be two candidate methods due to their potential advantages. Here we elaborate on them in detail.

Deep Learning: Traditional classification methods for HBR are only feasible for a few predefined behaviors. They require labeled data for supervised training. However, in practice, human behavior often exhibits unpredictability and high diversity, which make it impossible to label all behaviors. Furthermore, with the increase of mobile devices, mobile sensing data has become a new component of the big data community. A unified approach that enables management and analysis of these data captured from human daily behavior is still lacking. In view of these problems, seeking a general representation for all behaviors without prior knowledge is urgent for a deep and comprehensive understanding of mobile sensing data. According to the similarity measurement of the representation, raw sensing data of various behaviors can be segmented and classified automatically. We believe that such a deep learning scheme has considerable promise for wireless sensing applications such as localization and gesture recognition.

The research of human behavior perception has made remarkable achievements by using acoustic, optical, RF, and WiFi signals. Thus future research may tap more potential signals for behavioral recognition. Beyond amplitude and phase information in WiFi CSI, other information such as frequency and doppler shift will also be explored.

However, the deep learning idea also depends on analyzing the correlation between signal change and human behavior. In order to ensure the accuracy and robustness of the system, there are still challenges ahead.

Multi-path effect elimination. Wireless signals propagate in all radial directions, and reflect off various objects such as walls and furniture. Therefore, multiple copies of the same signal may reach a receiver with different delay and attenuation. This phenomenon is called the multi-path effect. To accurately recognize human behavior using CSI, it is essential to analyze the CSI changes caused only by human behavior. However, the raw CSI measurements are extremely noisy due to the multi-path effect. What is more, the multi-path effect fluctuates as the environment changes. Past solutions usually measure the multi-path without humans, and compare it to the signal when humans enter the environment. But when the environment changes, the previously measured no-human multi-path becomes outdated. Thus, how to design a highly robust mechanism to handle the complex multi-path effect is an enormous challenge.

Simultaneous perception of multiple body part movements. Prior works focus on either recognizing some large scale movements, such as walking, running, and sitting down, or merely sensing gestures of a certain body part, with the assumption that the other parts are static. However, in practice, people perform many activities by coordinating movements of multiple parts, rather than a single gesture. Thus, separating these movements is the basis for realizing behavior perception. But when multiple gestures occur simultaneously, the signal changes caused by them are mixed. How to distinguish the signal change caused by each body part movement is undoubtedly another challenging undertaking.

Signal Modeling: Many efforts will focus on establishing novel signal models by excavating and analyzing the physical nature of signal changes instead of training large amounts of template data. The key advantages of using signal modeling include that with zero human effort, it can sense an arbitrary set of body motions rather than a few predefined gestures, which enables wider ranges of real-time applications, especially for virtual handwriting application. Toward novel models between signal and human behavior, there is a tendency to further develop the theory for understanding issues such as how reflection, diffusion, and shadowing interfere with each other. The developed theory and model can also be applied to other applications.

Even if the models are feasible in theory, they still face significant challenges in applying them to real scenarios.

Dealing with the time asynchronous problem. Behavior recognition systems that utilize timestamps reported by commodity Wi-Fi cards can obtain time of flight (TOF) at a granularity of several nanoseconds, leading to tracking error of a few meters. Although some super-resolution algorithms are applied to obtain finer TOF estimates, the underlying assumption is that all APs (including transmitters and receivers) are time synchronized. That is almost impossible using commodity Wi-Fi deployments. Dealing with the time asynchronous problem in commodity Wi-Fi is a crucial as well as challenging issue in Wi-Fi CSI-based HBR.

Tiny reflected signal extraction. Some model based interfaces estimate the AOA of moving body parts by leveraging the reflected signal caused by human body. Nevertheless, to our best knowledge, the wireless signal arriving at receiver is a mixed signal from direct path and various reflected paths. Unfortunately, the human reflected signal is quite subtle compared to the signal from direct path and strong reflectors. Current solutions generally suppress undesired paths by prior measurement, while they are still unable to thoroughly attain the desired reflected component. Thus, it is very difficult to extract the tiny human reflected component from the mixed signal. Maybe the blind-source separation algorithm can serve as a solution.

Fresnel zone boundaries detection. In Fresnel zone model Wi-Fi-based HBR, the key is to quantitatively analyze how static and moving body parts affect the wireless signal in the Fresnel zone. Since the Fresnel zone area of a single link is very small, an interface usually employs multiple Wi-Fi devices to enlarge the coverage. The question is then what are the best, good, and bad locations and orientations of transceivers for behavior sensing. However, the Fresnel zones of multiple links are exceedingly complex, and how to detect the multiple Fresnel zone boundaries is a great challenge.

Conclusion

This article is mainly geared toward newcomers in the field of HBR using Wi-Fi CSI. We first briefly summarize the related work of HBR. We specifically focus on HBR using Wi-Fi CSI. We have reviewed recent progress of Wi-Fi CSI-based HBR, and divided them into two categories: fingerprint-based and model-based. Based on this classification, we describe the feature and general architecture of the two categories, and provide performance comparison between several existing approaches. Furthermore, we propose constructive insights about the future trends and research opportunities based on existing works and practical requirements. We hope that this tutorial proves helpful to encourage newcomers to design more advanced behavior recognition systems.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 61572402, No. 61672427, No. 61602382, and No. 61672428) and the International Cooperation Foundation of Shaanxi Province, China (No. 2015KW-003). Xiaoiang Chen and Yao Peng are corresponding authors.

References

[1] C. Liu *et al.*, "Lasagna: Towards Deep Hierarchical Understanding and Searching over Mobile Sensing Data," *Proc. MobiCom*, 2016, pp. 334–47.

[2] T. Liu *et al.*, "TagBooth: Deep Shopping Data Acquisition Powered by RFID Tags," *Proc. IEEE INFOCOM*, 2015, pp. 1670–78.

[3] X-box Kinect, Microsoft [EB/OL]; http://www.xbox.com, 2015.

[4] Leap motion, Leap Motion Corp., [EB/OL], https://www.leapmotion.com, 2015.

[5] W. Wang, A. X. Liu, and K. Sun, "Device-Free Gesture Tracking Using Acoustic Signals," *Proc. MobiCom*, 2016, pp. 82–94.

[6] W. Mao, J. He, and L. Qiu, "CAT: High-Precision Acoustic Motion Tracking," *Proc. MobiCom*, 2016, pp. 69–81.

[7] F. Adib and D. Katabi. "See through Walls with Wi-Fi," *Proc. SIGCOMM*, 2013, pp. 75–86.

[8] Q. Pu *et al.*, "Whole Home Gesture Recognition Using Wireless Signals," *Proc. MobiCom*, 2013, pp. 27–38.

[9] F. Adib *et al.*, "3D Tracking via Body Radio Reflections," *Proc. Usenix NSDI*, 2014, pp. 317–29.

[10] H. Wang *et al.*, "Human Respiration Detection with Commodity Wi-Fi Devices: Do User Location and Body Orientation Matter?," *Proc. IEEE INFOCOM*, 2016, pp. 25–36.

[11] W. Wang *et al.*, "Understanding and Modeling of Wi-Fi Signal Based Human Activity Recognition," *Proc. MobiCom*, 2015, pp. 65–76.

[12] Chunmei Han *et al.*, "WiFall: Device-free Fall Detection by Wireless Networks," *Proc. IEEE INFOCOM*, 2014, pp. 271–79.

[13] G. Wang *et al.*, "We can Hear you with Wi-Fi!," *Proc. MobiCom*, 2014, pp. 593–604.

[14] L. Sun *et al.*, "WiDraw: Enabling Hands-free Drawing in the Air on Commodity Wi-Fi Devices," *Proc. MobiCom*, 2015, pp. 77–89.

[15] H. Abdelnasser, M. Youssef and K. A. Harras, "WiGest: A Ubiquitous Wi-Fi-based Gesture Recognition System," *Proc. IEEE INFOCOM*, 2015, pp.1472–80.

Biographies

Lili Chen is a graduate student in the School of Information Science and Technology, Northwest University, Xi'an, China. Her current research interests include wireless signal localization and gesture recognition

Xiaojiang Chen received his Ph.D. degree in computer software and theory from Northwest University in 2010. He is currently a professor with the School of Information Science and Technology, Northwest University. His current research interests include localization and performance issues in wireless ad hoc, mesh, and sensor networks and named data networks.

Ligang Ni is a graduate student in the School of Information Science and Technology, Northwest University. His current research interests include wireless sensor networks and localization.

Yao Peng is a lecturer in the School of Information Science and Technology, Northwest University. Her current research interests include wireless sensor networks and localization. She mainly does research on wireless sensor networks.

Dingyi Fang received his Ph.D. degree in computer application technology from Northwestern Polytechnical University, Xi'an, China, in 2001. He is currently a professor with the School of Information Science and Technology, Northwest University. His current research interests include mobile computing and distributed computing systems, network and information security, and wireless sensor networks.

MOBILE BANDWIDTH IMPROVEMENT TECHNIQUES

Vijay K. Gurbani

Salvatore Loreto

Ravi Subramanyan

Wireless communication continues to grow, and with limited available spectrum and an ever increasing range of applications demanding access to that limited spectrum, techniques to improve throughput and reliability within existing constraints are being explored from many angles. These methods range from approaches that manage radio transmission (physical domain) such as mulitple-input multiple-output and interference alignment, to algorithmic and signal processing techniques, such as different coding methods.

The two articles in this issue of the Design and Implementation Series both discuss aspects of methods used for achieving improved bandwidth over the wireless medium. In the first article, "Interference Alignment Testbeds," Yetis *et al.* study the experimental evaluation of interference alignment (IA) in order to better quantify the limitations of existing, mostly analytical, results for IA. They point out that while much of the theoretical research on IA requires experimental evaluation to test under realistic conditions, available techniques for testing the efficacy of IA have been limited to simple configurations. Their article provides an overview of testbed implementations and discusses requirements for successful IA testing. The article also discusses required characteristics for successful future applications of IA to next generation wireless technologies.

Turning from the physical (radio) world, which is application-independent, to the specific application of LTE, the article "Practical LTE and Wi-Fi Coexistence Techniques beyond LBT" by Ling *et al.* discusses challenges for LTE to coexist with WiFi (i.e., for LTE to use additional bandwidth from unlicensed spectrum). The inability to detect collisions due to transmissions below the energy detection threshold makes it not possible for uLTE and WiFi to coexist and share spectrum smoothly. Their article uses simulations to show improvement in throughput when uLTE and Wi-Fi adapt their ED thresholds and coordinate, although their proposed technique does require reprogramming of existing WiFi access points.

BIOGRAPHIES

VIJAY K. GURBANI [M'98] (vijay.gurbani@nokia-bell-labs.com) is a Distinguished Member of Technical Staff at Bell Laboratories' End-to-End Mobile Network Research department in Nokia Networks. He holds a B.Sc. in computer science with a minor in mathematics and an M.Sc. in computer science, both from Bradley University; and a Ph.D. in computer science from Illinois Institute of Technology. His current work is focused on scalable analytic architectures and algorithms for autonomic 5G networks. His research has resulted in products that are used in national and international service provider networks. He has over 60 publications in peer-reviewed conferences and journals, five books, seven granted U.S. patents, and 19 IETF RFCs.

SALVATORE LORETO [M'01, SM'09] (salvatore.loreto@ieee.org) works as a strategic product manager within the Media business unit at Ericsson, Stockholm, Sweden. He has made contributions in Internet transport protocols (e.g., TCP, SCTP), signal protocols (e.g., SIP, XMPP), VoIP, IP-telephony convergence, conferencing over IP, 3GPP IP Multimedia Subsystem (IMS), HTTP, WebRTC, and web technologies. He is also a very active contributor to the IETF, where he has co-authored several RFCs and has served as co-chair for several working groups. For the IEEE Communications Society, he serves as a Design and Implementation Series co-Editor and Associate Technical Editor for *IEEE Communications Magazine*. He received an M.S. degree in engineer computer science and a Ph.D. degree in computer networking from Napoli University in 1999 and 2006, respectively. In 2014 he graduated as an executive M.B.A. from SDA Bocconi in Italy.

RAVI SUBRAHMANYAN [SM '97] (ravi.subrahmanyan@ieee.org) received M.S. and Ph.D. degrees in electrical engineering from Duke University, a B.Tech. from IIT Bombay, and an M.B.A. from MIT. He has over 50 refereed journal articles and conference publications, and holds over 20 issued patents. He has worked on various aspects of telecommunications, including hardware design and system architectures for data and video transport. He is a synchronization expert, was an Editor for *IEEE Communications Magazine* Feature Topics on Synchronization in NG Networks and NG911, and was a presenter on the Comsoc Webinar on Next Gen Synchronization Networks. He has served on various IEEE GLOBECOM and ICC conference committees and on ComSoc's TAOS TC since 2008, and is a Technical Editor for *IEEE Communications Magazine*.

Call for Papers

IEEE Communications Magazine

Multi-Access Mobile Edge Computing for Heterogeneous IoT

Background

The convergence of mobile internet and wireless systems have witnessed an explosive growth in resource-hungry and computation-intensive services and applications, which cover broad paradigms of so-called heterogeneous Internet of Things (H-IoTs). These systems include real-time video/audio surveillance, remote e-health systems, intelligent transportation systems, and Internet of Vehicles (IoV), and etc. Mobile edge computing, by placing various cloud resources (e.g., computational and storage resources) closer to smart devices/objects, has been envisioned as an enabling and highly promising technology to realize and reap the promising benefits of H-IoTs applications. However, the growing demands for ultra-low latency, massive connectivity, and high reliability of the large number of H-IoTs applications has yielded a critical issue in mobile edge computing, i.e. the limited connections (such as connection capacity, bandwidth, or the number of simultaneously affordable connections) between mobile edge cloud and smart devices/objects.

Multi-access mobile edge computing (MA-MEC), which actively exploits a systematic and adaptive integration of recent radio access technologies including 5G, LTE, and Wi-Fi to enhance the access capacity of smart devices to mobile edge platforms, has been considered as a highly promising technology to tackle this issue. The evolution towards the architecture of ultra-dense small-cells (micro / pico / femto cells, and Wi-Fi hotspots) in future radio access networks facilitates the MA-MEC, i.e., the densely deployed small cells can significantly improve the capacity and quality of the connections between smart devices and mobile edge cloud. For instance, the emerging small-cell dual-connectivity in small-cell networks enables smart mobile devices to communicate with conventional macro-cells and off load data traffic to small cells simultaneously. This enhances the access capacity of mobile edge cloud at small cells.

Therefore, with the strength of multi-access for capacity-enhancement, the MA-MEC is expected to bring a variety of benefits, such as i) ultra-low latency between smart devices and edge cloud for real-time, interactive, and mission-critical applications, e.g., the real-time indoor navigation and augmented virtual-reality, ii) privacy and security in local communications to access mobile edge cloud, and iii) the big data analytics at the point of capture for IoT applications. For instance, the MA-MEC can facilitate the implementation of various safety-oriented applications in transport systems, in which MA-MEC provides robust and ultra-low latency connections for smart vehicles to efficiently access mobile edges for real-time safety-related information processing at mobile edge at the road-side units.

However, the success of MA-MEC still requires tackling many new challenges. To efficiently exploit computation and storage resources at mobile edge nodes, a joint optimization of placement of computation/storage resource and cell-association with radio resource allocation is necessitated. Such joint optimization should be adaptive according to time-varying environments, e.g., the varying wireless channel states when users move across the cells and the dynamic computation/storage resource utilizations. Therefore, this Feature Topic (FT) aims at soliciting high quality and unpublished work regarding recent advances in MA-MEC, with the main focus on addressing the fundamental design issues in MA-MEC, and the emerging paradigms and testbeds that use MA-MEC. We solicit papers covering the topics of interests in the following two main categories:

- Fundamental design issues in MA-MEC
- Radio resource management for MA-MEC
- Task scheduling and computation resource management for MA-MEC
- Virtualization and network slicing for MA-MEC
- Location and sizing of computation and storage elements for MA-MEC
- Communication protocols and network architectures for MA-MEC
- Security, privacy, and reliability in MA-MEC
- QoE and QoS provisioning in MA-MEC
- 5G/LTE/WiFi enabled MA-MEC
- Energy management and green MA-MEC
- Edge-to-cloud integration and protocols for MA-MEC
- Human and social-driven design of MA-MEC
- MA-MEC for Heterogeneous IoT
- MA-MEC for smart cities
- MA-MEC for video/audio surveillance
- MA-MEC for industrial IoT
- MA-MEC for smart energy systems
- MA-MEC for smart healthcare
- MA-MEC for intelligent transportation systems
- MA-MEC for big data analytics

Submissions

Articles should be tutorial in nature, with the intended audience being all members of the communications technology community. They should be written in a tutorial 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, excluding figures, tables and captions). Figures and tables should be limited to a combined total of six. The number of archival 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 submit 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 "July 2018 / Multi-Access Mobile Edge Computing for Heterogeneous IoT" as the Feature Topic category for your submission.

Important Dates

- Manuscript Submissions Deadline: November 1, 2017
- Decision Notification: March 1, 2018
- Final Manuscripts Due: April 15, 2018
- Publication Date: July 2018

Guest Editors

Yan Zhang
University of Oslo, Norway
yanzhang@ieee.org

Yuan Wu
Zhejiang Univ. of Technol., China
iewuy@zjut.edu.cn

Hassnaa Moustafa
Intel Corporation, USA
hassnaa.moustafa@intel.com

Danny H.K. Tsang
HKUST, Hong Kong
eetsang@ust.hk

Alberto Leon-Garcia
Univ. of Toronto, Canada
alberto.leongarcia@utoronto.ca

Usman Javaid
Vodafone, UK
usman.javaid@vodafone.com

Interference Alignment Testbeds

Cenk M. Yetis, Jacobo Fanjul, José A. García-Naya, Nima N. Moghadam, and Hamed Farhadi

The authors summarize the practical limitations of experimentally evaluating IA, provide an overview of the available IA testbed implementations, including the costs, and highlight the imperatives for the succeeding IA testbed implementations. They also explore future research directions on the applications of IA in the next generation wireless systems.

ABSTRACT

Interference alignment has triggered high impact research in wireless communications since it was proposed nearly 10 years ago. However, the vast majority of research is centered on the theory of interference alignment and is hardly feasible in view of the existing state-of-the-art wireless technologies. Although several research groups have assessed the feasibility of interference alignment via testbed measurements in realistic environments, the experimental evaluation of interference alignment is still in its infancy since most of the experiments were limited to simpler scenarios and configurations. This article summarizes the practical limitations of experimentally evaluating interference alignment, provides an overview of the available interference alignment testbed implementations, including the costs, and highlights the imperatives for succeeding interference alignment testbed implementations. Finally, the article explores future research directions on the applications of interference alignment in the next generation wireless systems.

INTRODUCTION

The recently developed interference alignment (IA) concept has revealed that the throughput of a wireless network can be significantly improved compared to that exhibited by conventional transmission schemes such as time-division multiple access (TDMA) and frequency-division multiple access (FDMA). Unfortunately, it is extremely difficult to take into account all practical limitations in the analytical investigation of IA, yielding theoretical results that are frequently based on assumptions which are hardly realizable in real-world scenarios. Examples of such practical aspects are the impacts of imperfect channel state information (CSI), energy loss due to spatial collinearity between desired signal and interference subspaces, detection and synchronization errors, and imperfect hardware. Consequently, the experimental evaluation of IA techniques is crucial to better understand the impacts of the aforementioned practical limitations on the performance of existing IA techniques as well as to propose new research topics around the IA concept to overcome such limitations. This is precisely the main goal of this article, which consists of proposing future research directions aiming for adopting IA to an attractive solution to be considered by the industry for the next generations of wireless communication systems.

The original IA concept assumes perfect CSI knowledge in all terminals to design the corresponding beamformers and filters. In practice, however, users can acquire only a noisy version of the CSI, yielding a significant performance degradation in terms of the achievable sum degrees of freedom (DoF), as shown in [1] for a pilot-assisted channel estimation technique in a *K*-user interference network with single-antenna users. The CSI acquisition problem is mitigated in time-division duplexing (TDD) systems by exploiting channel reciprocity, although calibrated RF equipment is required [2]. For frequency-division duplexing (FDD) systems, IA experiments with perfect [3, 4] and realistic analog wireless [5] feedback channels have been reported in the literature. Furthermore, in most of the theoretical IA works, the block-fading channel model assumption plays a pivotal role due to its mathematical tractability [1]. In practice, guaranteeing a constant channel during a block transmission is not possible, leading to the additional problem of outdated CSI at the transmitters.

The implementation of a perfect IA scheme requires a large number of transmitters (e.g., base stations in the downlink of cellular networks) and network resources (i.e., time, frequency, number of antennas, and power). Particularly, the number of signal space dimensions grows exponentially with the number of users. Hence, in a *K*-user interference network, *K* transmitters are required to serve *K* users. Therefore, further research is needed on multiple-input multiple-output (MIMO) interference relay broadcast channels to serve more users per transmitter via relays with confined resources.

The finite signal-to-noise ratio (SNR) at the receiver is another practical limitation. However, experimental results show that, despite the imperfections in both CSI acquisition and testbed hardware, IA outperforms conventional communication schemes such as TDMA and greedy interference avoidance in the mid-to-high SNR regime [4]. However, an optimal trade-off between network resources dedicated to CSI acquisition and feedback with respect to those devoted to data transmissions must be determined to maximize the throughput [1, 6, 7]. Error vector magnitude (EVM) experimental results for a pilot-assisted maximum signal-to-interference-plus-noise ratio (max-SINR) scheme corroborate the existence of an optimal resource allocation scheme [7] and an optimal number of training symbols [4].

Time and frequency synchronization between network nodes is of utmost importance when experimentally evaluating IA techniques, and it can be implemented in a centralized [8] or distributed manner [2, 5]. Theoretical IA works usually assume

Digital Object Identifier: 10.1109/MCOM.2017.1600471

Cenk M. Yetis is with Academia Sinica; Jacobo Fanjul is with the University of Cantabria; José A. García-Naya is with the University of A Coruña; Nima N. Moghadam is with KTH Royal Institute of Technology; Hamed Farhadi is with Harvard University.

0163-6804/17/$25.00 © 2017 IEEE

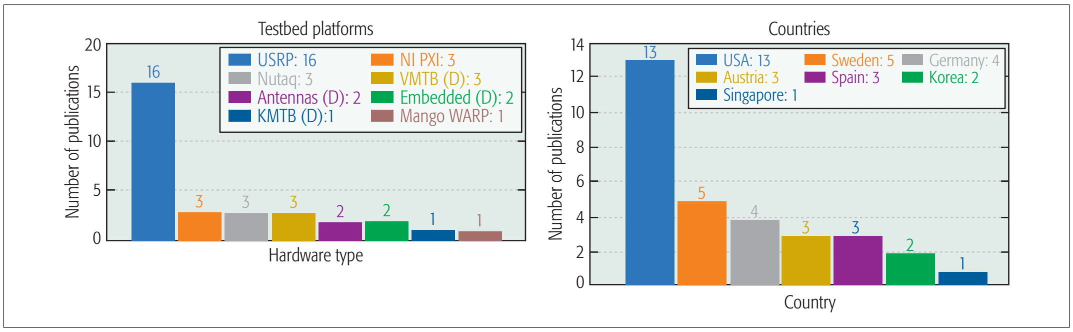


Figure 1. From left to right, the number of IA testbed publications (only officially published papers are considered) per hardware type and per country are given, respectively. D: partially dedicated testbed platforms, where there are specifically designed hardware components along with off-the-shelf products; KMTB: Kista MIMO testbed in Sweden; VMTB: Vienna MIMO testbed in Austria; antennas: testbed implementations with antennas that have different radiation patterns; embedded: embedded implementations.

that beamformers operate after frame detection and synchronization. In practical systems, however, frame detection and synchronization are applied immediately after the analog-to-digital conversion, hence being affected by interference and yielding a strong performance degradation of such tasks, thus impacting dramatically on the final system performance. For the specific case of spatial IA in multicarrier systems, IA decoding can be implemented in the time domain or following a more conventional per-subcarrier approach in the frequency domain [3, 4]. Given that time-domain IA decoding suppresses most of the multiuser interference at the very beginning of the receiver signal processing chain, the effective SINR is improved, whereas synchronization tasks perform similar to the interference-free case because they operate after filtering out the interference [4].

Hardware imperfections are ignored in many IA algorithm designs. However, nonlinear distortions, phase noise, IQ imbalance, and frequency offset degrade IA performance [3, 9]. For instance, the measurement results of IA in the 3-user 2 × 2 MIMO interference channel show that hardware imperfections can reduce the maximum achievable SINR up to 10 dB compared to the theoretical predictions [9]. To compensate for the leaked interference under non-ideal conditions, power control is suggested as a complementary interference management technique [10].

This article addresses the main practical limitations that have been found when experimentally evaluating IA, and describes solutions to mitigate their impacts on the IA performance. The rest of the article is structured as follows. First, a panorama of the testbeds that have been employed to validate different IA techniques is provided. An estimation of the cost of these testbeds and statistics regarding the publications associated to them are provided. Finally, the article explores future directions to be taken by both theoretical and experimental IA investigations for enabling IA in the next generation of wireless communications.

Highlights of Testbed Implementations

In this section we provide statistical information on IA experiments and publications, as well as financial costs of these IA implementations. Various options that span from low- to high-end solutions are also summarized.

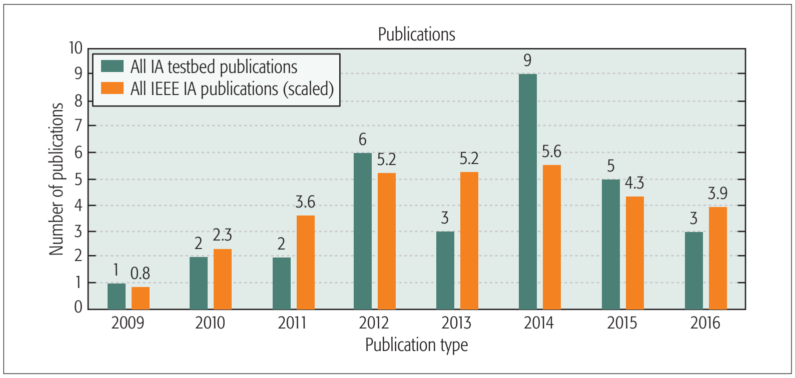


Figure 2. All IA testbed publications (including non-IEEE publications) and the scaled number of all IEEE IA publications (including non-testbed publications) per year. The scaling factor 47 is obtained from the ratios of the averages of the two datasets per year.

Publications on IA Experiments

Some statistics regarding the number of IA publications are shown in Figs. 1 and 2. As the IA concept evolves, testbeds incorporate a mixture of dedicated hardware components and commercial off-the-shelf modules (we refer to them as dedicated platforms). When the concept matures, the amount of dedicated hardware components in the testbed platforms overrun. As shown in Fig. 1, commercial off-the-shelf products are still the leading choices for IA implementations within the research community. On the other hand, as seen in Fig. 2, the number of IA testbed publications fluctuates from year to year, and each year a fairly good number of experimental studies are published.

Analysis of the Costs of Equipment in IA Experiments

In Table 1, the equipment used in each of the reported IA testbeds and its estimated costs are listed. As can be seen in the list, a very affordable IA setup is demonstrated in [8] where transmit antenna selection is applied, and two antennas out of three are selected. The next affordable setup is reported in [5]. With a similar total cost, high-performance universal software radio peripherals (USRPs) support 2 × 2 MIMO configurations, as shown in Table 1. The first IA real-time implementation is introduced in [11], where blind interference alignment (BIA) is implemented for a 2-user 2 × 1 broadcast channel. In this setup, Commercial off-the-shelf products are still the leading choices for IA implementations within the research community. On the other hand, the number of IA testbed publications fluctuates from year to year, and each year a fairly good number of experimental studies are published.

there are two antennas at the transmitter, whereas at the receiver, one of two antennas is selected. Except for [8, 11], all configurations in Table 1 are outlined for the 3-user 2 × 2 interference channel with a single stream per user, $(2 \times 2,1)^3$, including the equipment used in the recent implementation of a massive MIMO testbed by the University of Bristol and Lund University collaboration. With a comparable total cost and testbed setup, a similar equipment list is used in the centralized PXI configuration. Finally, in the distributed PXI configuration, a separate chassis is used for each transmitter.

In Fig. 3, the estimated costs of selected configurations from Table 1:

- Configuration C: high performance USRPs
- Configuration F: centralized PXI
- Configuration G: distributed PXI

are plotted vs. the network size.

In the massive MIMO demonstration by the universities in Bristol and Lund, 16 users are served by a 128-antenna transmitter. Accordingly, in order to implement IA in the $(9 \times 8,1)^{16}$ scenario, by using the same equipment listed in Table 1 for configurations F and G, a total cost of nearly US$2 million is estimated. As shown in Fig. 3, to scale the network DoF by a factor of four, that is, from 3: $(2 \times 2,1)^3$ to 12: $(8 \times 7,3)^4$, the cost is expected to increase nearly four times, whereas by scaling the DoF approximately by a factor of five, that is, from 3 to 16: $(9 \times 8,1)^{16}$, the cost is expected to scale by a factor of 20. Hence, for large network sizes where IA is significantly competitive, large capital investments are required.

IA Testbed Platforms and Experiments

The number of solutions devoted to experimental research in wireless communications is growing every year, particularly with the increased interest in wireless sensor networks (WSNs) and the Internet of Things (IoT) from the research community. Publicly accessible testbed facilities are great opportunities for IA researchers to experimentally evaluate their algorithms. However, among the large-scale testbed facilities that offer public access for researchers to execute automated and manageable experiments, only a few of them are suitable for physical-layer experimentation. A good example is CorteXlab at the University of Lyon, which contains a mixture

Configuration type		[8]	[5]	High-perf. USRP	[11]	Bristol & Lund	Cent. PXI	Dist. PXI
		A	B	C	D	E	F	G
Equipment with model number and part number	Unit cost* of equipment	Quantity of equipment used in a configuration						
		A	B	C	D	E	F	G
OctoClock CDA-2990 782978-01	1.1	1	–	1	-	1	1	1
NI USRP-2943R 783925-01	6.8	-	-	6	-	3	3	3
NI USRP-2953R, GPS Clock 783928-01	8.1	-	-	-	-	3	3	3
NI PXIe-1082 Chassis 780321-01	3.8	-	-	-	3	-	-	3
NI PXIe-1085 Chassis 783588-01	10.0	-	-	-	-	1	1	-
NI PXIe-7976R FlexRIO FPGA 783625-01	11.0	-	-	-	-	3	3	3
NI PXIe-8840 RT Controller 783001-33	5.1	-	-	-	-	1	-	3
NI PXIe-8880 Controller 783513-33	8.0	-	-	-	-	-	1	-
USRP N210 782747-01	2.1	6	-	-	-	-	-	-
USRP B210 782981-01/784190-01	1.4	3	-	-	-	-	-	-
NI USRP-2921 781907-01	2.8	-	12	-	-	-	-	-
GPSDO Kit for USRP N200/N210 782779-01	0.9	-	6	-	-	-	-	-
NI PXIe-8130 Controller	5.1	-	-	-	3	-	-	-
NI PXIe-7965R FlexRIO FPGA 781207-01	10.1	-	-	-	4	-	-	-
NI 5781 Baseband Transceiver 781267-01	3.3	-	-	-	4	-	-	-
XCVR2450	0.5	-	-	-	4	-	-	-
Total Cost* of a Configuration (* Cost in thousand USD)		18	40	42	82	94	97	105

Table 1 . The quantity needed per equipment and the total estimated cost are listed for exemplary IA testbeds and for some possible IA testbed configurations. For simplicity, only the main equipment is listed; software and other equipment costs, including cables, antennas, and PCs, are not listed in the table. The rounded prices are in thousands of U.S. dollars and obtained from the National Instruments US website.

of low-power, general-purpose, and real-time high-performance nodes. Other small-scale publicly available facilities are CREW and CORE+ project consortiums.

There are different platforms available for the implementation of an IA testbed. National Instruments and Ettus USRPs together with the open source universal hardware driver and GNU Radio are among the preferred choices for low-budget cases. Open source implementations of communication standards like openLTE for the case of LTE and the enormous user community make this solution very attractive. At an increased cost, National Instruments USRPs can be preferred since they support LabView, a powerful software that is another proprietary product of National Instruments. Other examples of low-cost solutions available in the market are RTL-SDR, HackRF, and Nuand BladeRF. Unfortunately, their simple designs and limited capabilities make complex IA implementations on these platforms infeasible.

When the budget is not a severe constraint, high-end PXI-based products are much better solutions. There are several PXI hardware vendors, such as Keysight and National Instruments, where the main advantage of National Instruments solutions lies on the software side, especially their integration with LabView and with other powerful software tools like MATLAB from Mathworks.

Many manufacturers offer hardware equipment besides the integrated solutions, allowing for developing a part-by-part testbed. 4DSP, Nutaq, and Innovative Integration are just some examples. Some of the manufacturers offer modules for Xilinx-based boards, and typically they do not provide full solutions or open source drivers. At a lower level, Texas Instruments, Maxim Integrated, and Analog Devices are among the manufacturers continuously offering better components for SDR solutions. For example, Analog Devices has introduced a transceiver (AD-FMCOMMS5-EBZ) with up to 8 antennas and 56 MHz bandwidth on a single board. At a much higher level, the Wireless Open Access Research Platform (WARP) is an example of a full bundle of solutions built from the ground up with the aim of prototyping advanced wireless networks. Several alternatives similar to WARP can be found within the wireless research community, for example, the SDR4All project from Supélec.

Frequently, research institutions opt for developing their testbeds based on commercial off-the-shelf hardware, sometimes combined with custom-designed parts. Examples of such testbeds are those developed by the Vodafone Chair for Mobile Communications Systems at Techsnische Universitt Dresden, KTH Testbed, OpenAirInterface at Eurecom, the so-called Vienna MIMO Testbed developed at the Institute of Telecommunications at Technische Universität Wien, and the one developed at the Heinrich Hertz Institute at the Fraunhofer Institute for Telecommunications.

Once the hardware required to experimentally evaluate IA techniques is ready, a challenge that arises is the generation of a representative and sufficiently large amount of channel realizations guaranteeing repetitive and reproducible results. Fortunately, many clever and inspiring approaches are found in the literature, for example, considering antenna switching instead of reconfigurable antennas [11], utilizing two different reconfigurable antenna architectures that use different patterns, or simply sliding the receive antenna.

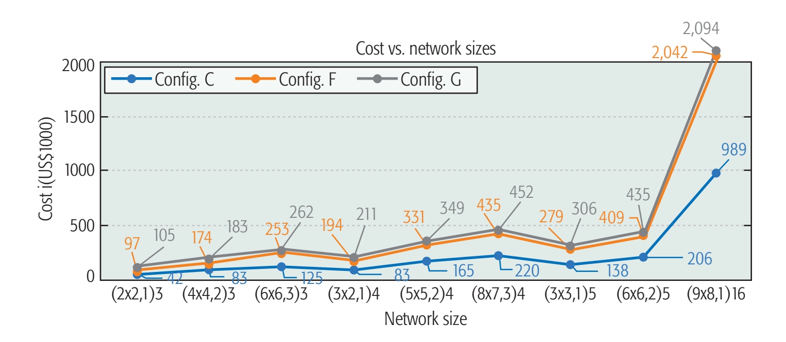


Figure 3. Costs, expressed in thousands of U.S. dollars, of selected configurations from Table 1 vs. network sizes.

The aforementioned experiments rely on physical techniques (i.e., different antenna architectures) to supply independent channel realizations. However, independent channel realizations can also be created by simply conjugating the complex-valued input and output signals. Thus, IA can be achieved via the conjugate operation in static single-input single-output (SISO) X networks without any DoF loss, and in interference networks with some DoF loss.

Future Requirements

In this section we provide an overview on the future requirements to transform IA into an attractive solution to be considered for the next generations of wireless communication systems.

Scalability and Measurements

As mentioned before, existing experimental setups cover simplified scenarios with a reduced number of nodes and antennas per node. However, real-world wireless networks usually include a large number of users and base stations equipped with several antennas each. Even though the implementation of sophisticated nodes is much more expensive, this aspect is of utmost importance regarding the evaluation of IA in realistic scenarios. Experimental evaluation of wireless communication systems in general requires certain measurement concepts and techniques such as the treatment of uncertainties in the results [12]. As the network size scales, the application of this discipline becomes more complex and also more vital, and hence such measurement concepts and techniques need to be adapted to the particular case of IA as well.

Different Network Topologies

The IA testbed implementations summarized in the previous sections can be gathered under two groups: IA with channel state information at the transmitters (CSIT) in interference networks and IA with no CSIT, for example, BIA in broadcast networks.

Another promising direction is the implementation of IA in relay networks, either with or without CSIT. For the former, innovative techniques are emerging, such as aligned network coding and

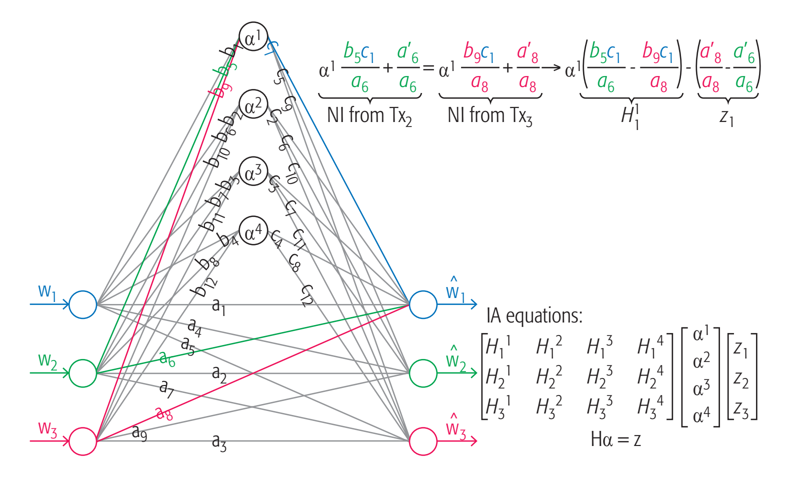

$\alpha^1 \frac{b_5c_1}{a_6} + \frac{a'_6}{a_6} = \alpha^1 \frac{b_9c_1}{a_8} + \frac{a'_8}{a_8} \rightarrow \alpha^1\left(\frac{b_5c_1}{a_6} - \frac{b_9c_1}{a_8}\right) - \left(\frac{a'_8}{a_8} - \frac{a'_6}{a_6}\right)$
NI from Tx_2
NI from Tx_3
H_1^1
z_1
α^1
α^2
α^3
α^4
b_1 b_5 b_9 b_2 b_6 b_{10} b_3 b_7 b_{11} b_4 b_8 b_{12}
c_1 c_5 c_9 c_2 c_6 c_{10} c_3 c_7 c_{11} c_4 c_8 c_{12}
a_1 a_4 a_5 a_6 a_2 a_7 a_8 a_9 a_3
w_1
w_2
w_3
$\hat{w}_1$
$\hat{w}_2$
$\hat{w}_3$
IA equations:
$\begin{bmatrix} H_1^1 & H_1^2 & H_1^3 & H_1^4 \\ H_2^1 & H_2^2 & H_2^3 & H_2^4 \\ H_3^1 & H_3^2 & H_3^3 & H_3^4 \end{bmatrix} \begin{bmatrix} \alpha^1 \\ \alpha^2 \\ \alpha^3 \\ \alpha^4 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$
$H\alpha = z$

Figure 4. 3-user relay-aided SISO-IC. α^q is the scaling factor of the qth relay. a_i and a_i' are the direct channels in the first and second time slots, respectively. b_i and c_i are the channels between the transmitters and relays, and the relay and receivers, respectively. The indirect channels are subindexed so that the desired and interference signals are carried over the effective channels b_ic_i and b_ic_j, $j \neq i$ respectively. H_i^q is the difference of the normalized interferences received via relays, $\forall i = \{1, 2, ..., K\}$ and $\forall q = \{1, 2, ..., Q\}$ where Q is the total number of relays in the system. z_i is the difference of the normalized interferences received via direct channels in the second time slot.

aligned interference neutralization. For the latter, BIA in relay networks is particularly appealing since IA drawbacks in broadcast or X networks are eliminated via the relay nodes. For conventional SISO interference channels (SISO-ICs) without relays, time-varying channels with long symbol extensions are required to obtain the optimal DoF. This requires an overwhelming amount of channel feedback overhead to each transmitter. However, in the relay-aided SISO-IC, only two time slots are required, and hence the feedback overhead is greatly reduced. Moreover, the relays are located in between the transmitters and the receivers, and thus have more accurate CSI feedback from the receivers compared to the CSI feedback to the transmitters, which are located farther from the receivers. Hence, the relay-aided BIA schemes are appealing from a practical perspective. Finally, as mentioned above, extending the BIA schemes to interference broadcast relay channels is also important to serve more users with fewer base stations deployed in a cellular network.

Nevertheless, regarding the next-generation communications, more complex and advanced network topologies, such as heterogeneous networks (HetNets), should also be considered. Some of the HetNet scenarios seem to be addressable in terms of IA testbed implementation. For instance, a reverse TDD (R-TDD) scenario could be deployed using relatively small and simple nodes. A 2-cell R-TDD setting with these characteristics would be challenging, especially in the CSI feedback aspect, but a combination of IA with non-coherent approaches (e.g., Grassmannian signaling) could also be considered in order to overcome this issue. Altogether, a successful implementation of IA techniques for these kinds of topologies would be a significant step forward in terms of the feasibility of alignment-based transmissions for the next-generation wireless communications.

The Multi-Stream Milestone

Multi-stream transmissions are essential in modern wireless communication systems, for example, for the transmission of rich media. However, they have to consider more aspects compared to their single-stream counterpart. Two good examples are the increased number of quality of service (QoS) parameters, such as multi-bit-rate streaming, and self-interference due to multiple information streams for each user. Unfortunately, very few IA experimental evaluations have been carried out considering multi-stream scenarios (e.g., [13]). One reason is that the implementation cost of multi-stream transmission is folded, as illustrated in Fig. 3. Moreover, to properly address the experimental evaluation of the throughput of a multi-stream (and also a single-stream) scenario, a minimal MAC layer implementation is necessary. A different approach would be to properly evaluate the performance of the physical layer and later plug in the results in a network emulator that can transmit real-world data.

Outdoors, High Mobility, and Reverberation Chambers

Few IA testbed implementations have considered outdoor and/or mobile environments, whereas the majority have addressed only static indoor scenarios. To the authors' knowledge, there is only one IA testbed implementation that is close to a prototype stage since IA beams are physically transmitted over the air and the complete system is realized in real time [5]. Nowadays, testbed equipment can be powered by small batteries and controlled using a laptop (e.g., B-family USRPs), thus making it possible to assess IA in high-mobility scenarios in the near future [14].

Reverberation chambers provide repeatable emulations of reasonably realistic conditions at relatively affordable costs. Similar to the success story of the WARP platform, which was initially developed at Rice University and later gave way to the spin-off company Mango Communications, the OTA reverberation chamber was initiated at Chalmers University of Technology for research purposes, and then the technology was transferred to the spin-off company Bluetest AB.

Software Requirements

Until now, the discussions are typically dominated by hardware specifications. However, a common repository to share and improve open source IA software by the developers is perhaps another immediate need. Such an approach can provide huge momentum for in-depth research and for the expansion of IA applications.

Reflections on Future Directions

As stated above, IA offers two main routes in theory: IA with CSIT, and IA with no CSIT or, briefly, BIA. While both IA schemes still face the perfect synchronization and imperfect hardware hurdles, BIA, especially in relay networks, is promising since the drawbacks of CSIT and unlimited network resources are eliminated via relay nodes. However, the high SNR requirement, another major IA drawback, remains in both IA schemes. Many complex schemes are proposed to improve the mid-SNR performance of IA at the cost of impairing its simplicity. Relay nodes can easily be utilized to improve the mid-SNR performance of IA as well. In relay networks, BIA is presently known to be more compatible than IA with CSIT.

Promising Directions

Among the options of the relay-aided SISO-IC scheme [15], discussed further in the next section, the option of multiple single-antenna relays in which each relay has a single antenna is less favorable since it has two requirements: joint beamforming between transmitters and relays, and time-varying channels. On the other hand, the option of a single relay with multiple antennas is more favorable since it does not impose those requirements. However, this option can only be favorable for mid-sized networks; for example, in a 21-user relay-aided SISO-IC, a single relay with 20 antennas can achieve IA. However, for a larger network, for example, in a 100-user relay-aided SISO-IC, IA can be achieved either via a single relay with 99 antennas or via 25 relays with 20 antennas each. Relay-aided MIMO-ICs can also be preferred for mid-sized networks since multiple antennas at the transmitters and receivers increase the DoF, but also increase the number of antennas at the relay. When each of the transmitters and receivers has two antennas, a single relay with 20 antennas can achieve IA in an 11-user MIMO-IC. However, 30 antennas are needed at the relay when each of the end nodes has three antennas.

IA techniques are currently difficult to implement in cellular networks, at least not in the medium term. Hence, massive MIMO, HetNets, and even IoT in cellular networks seem to be unpromising as immediate industrial pursuits for IA. In massive MIMO networks, the massive amount of nodes and antennas imply massive CSI needs. BIA can be implemented in such networks where massive MIMO nodes are the relay nodes in the next generation of cellular systems right after the launch of massive MIMO technology. Other possible technology options to choose from can be IoT in a future Bluetooth release or in WiFi HaLow networks, green networks, energy harvesting, Li-Fi, body area networks (BANs), and sensor networks. As a particular application, relay-aided BIA can be fine-tuned to be utilized in homes or commercial airplanes (IoT), on persons (BANs), in solar panels, or in smart farms (sensor networks), in airplanes, trains, ships, or cars (sensor networks), and in server rooms (green networks) to connect and monitor communicating wireless devices. After choosing the most compatible application for relay-aided BIA, theoretical and experimental-based research must progress in parallel at the same pace so that frequent feedback between them expedites the marketization process of IA.

Relay-Aided IA Schemes

The relay-aided IA scheme utilizes the conventional IA concept along with simply counting the number of variables and equations in the system. The concept is illustrated for the three-user SISO interference relay channel in Fig. 4. Since it is a two-time-slot scheme, each receiver has a 2D space. Hence, $K - 1$ interfering signals need to be aligned in a 1D space. Without loss of generality, this can be achieved as follows. For receiver 1, the normalized interference (NI) from transmitter 2 can be equated to the NI from each of the other transmitters: NI from Tx$_2$ = NI from Tx$_j$, $\forall j = \{1, 2, ..., K\}\backslash\{1, 2\}$ at Rx$_1$. For each of the other receivers, the normalized interference from transmitter 1 can be equated to the normalized interference from each of the other transmitters: NI from Tx$_1$ = NI from Tx$_j$, $\forall j = \{1, 2, ... K\}\backslash\{1, i\}$ at Rx$_i$, $\forall i = \{1, 2, ..., K\}\backslash\{1\}$. This is illustrated in Fig. 4 via the colored channels for the communication to receiver 1 through relay 1 only. The IA equations of the system can be reformulated in a matrix structure $\mathbf{H}\alpha = \mathbf{z}$ as defined in Fig. 4. The core idea of the proposed scheme is that the z vector cannot be zero when the number of constraints (i.e., number of rows in **H**) is equal to the number of variables (i.e., number of columns in **H**). Otherwise, since **H** is invertible, the scaling factors of relays are all zero, $\alpha = \mathbf{0}$. Therefore, the transmitters during the second time slot cannot be silent (i.e., a_i's cannot be zero), and the channels must be time varying (i.e., a_i' cannot be equal to a_i). Since there are $K(K - 2)$ IA equations in total, there need to be $K(K - 2)$ relays in the system so that there are $K(K - 2)$ variables (i.e., scaling factors) as also proposed in [15]. However, as illustrated in Fig. 4, with the addition of a relay, **H** is not square; hence, **z** vector can be zero, meaning that joint beamforming (i.e., transmitters also transmit during the second time slot) and time-varying channels are not needed.

Conclusion

In this article, a survey on IA testbed implementations has been presented. Highlights, challenges, and future directions of IA experimentations are provided from a broad perspective. Testbed experiments on the feasibility and performance of IA schemes have demonstrated significant gains compared to more conventional schemes. However, IA experimentation is still in its infancy, leading to a big gap between theoretical and experimental progresses. Moreover, IA testbed platforms notably lack many components compared to other worldwide deployed testbeds that have sophisticated configurations and features. More collaborations with both computer scientists and engineers as well as with other specialists in electronics, including microelectronics, can expedite the delivery date of IA to real-life.

IA testbed platforms notably lack many components compared to other worldwide deployed testbeds that have sophisticated configurations and features. More collaborations with both computer scientists and engineers as well as with other specialists in electronics, including microelectronics, can expedite the delivery date of IA to real life.

Acknowledgment

Jacobo Fanjul's research has been supported by the Ministerio de Economíca y Competitividad (MINECO) of Spain, under grants TEC2013-47141-C4-R (RACHEL project) and FPI grant BES-2014- 069786. José A. García-Naya's research has been funded by the Xunta de Galicia (ED431C 2016-045, ED341D R2016/012, ED431G/01), the Agencia Estatal de Investigación of Spain (TEC2013-47141-C4-1-R, TEC2015-69648-REDC, TEC2016-75067-C4-1-R), and ERDF funds of the EU (AEI/FEDER, UE). Hamed Farhadi's research has been funded by the Swedish Research Council (VR) under grant 2015-00500.

References

[1] H. Farhadi, M. N. Khormuji, and M. Skoglund "Pilot-Assisted Ergodic Interference Alignment for Wireless Networks," *Proc. 2014 IEEE ICASSP*, 4–9 May 2014, pp. 6186–90.

[2] S. Gollakota, S. D. Perli, and D. Katabi, "Interference Alignment and Cancellation," *Proc. ACM SIGCOMM*, Aug. 17–21, 2009.

[3] C. Lameiro *et al.*, "Experimental Evaluation of Interference Alignment for Broadband WLAN Systems," *EURASIP J. Wireless Commun. and Networking*, vol. 2015, issue 180, June 2015.

[4] J. Fanjul *et al.*, "An Experimental Evaluation of Broadband Spatial IA for Uncoordinated MIMO-OFDM Systems," *Proc. IEEE Int'l. Conf. DSP*, 21-24 July 2015, pp. 570–74.

[5] S. Lee, A. Gerstlauer, and R. W. Heath, "Distributed Real-Time Implementation of Interference Alignment with Analog Feedback," *IEEE Trans. Vehic. Tech.*, vol. 64, no. 8, Aug. 2015, pp. 3513–25.

[6] O. El Ayach, A. Lozano, and R. W. Heath, "On the Overhead of Interference Alignment: Training, Feedback, and Cooperation," *IEEE Trans. Wireless Commun.*, vol. 11, no. 11, Nov. 2012, pp. 4192–4203.

[7] N. N. Moghadam, H. Farhadi, and P. Zetterberg, "Optimal Power Allocation for Pilot-Assisted Interference Alignment in MIMO Interference Networks: Test-Bed Results," *Proc. 2015 IEEE Int'l. Conf. DSP*, 21-24 July 2015, pp. 585–89.

[8] M. El-Absi *et al.*, "Antenna Selection for Reliable MIMO-OFDM Interference Alignment Systems: Measurement Based Evaluation," *IEEE Trans. Vehic. Tech.*, vol. 65, no. 5, May 2015, pp. 2965–77.

[9] P. Zetterberg and N. Moghadam, "An Experimental Investigation of SIMO, MIMO, Interference-Alignment (IA) and Coordinated Multi-Point (CoMP)," *Proc. IWSSIP*, 11–13 Apr. 2012, pp. 211–16.

[10] H. Farhadi, C. Wang, and M. Skoglund, "Distributed Transceiver Design and Power Control for Wireless MIMO Interference Networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, Mar. 2015, pp. 1199–1212.

[11] K. Miller *et al.*, "Enabling Real-Time Interference Alignment: Promises and Challenges," *Proc. ACM MobiHoc*, 11–14 June 2012, pp. 55–64.

[12] S. Caban, J. A. Garcia-Naya, and M. Rupp, "Measuring the Physical Layer Performance of Wireless Communication Systems: Part 33 in a Series of Tutorials on Instrumentation and Measurement," *IEEE Instrumentation and Measurement Mag.*, vol. 14, no. 5, Oct. 2011, pp. 8–17.

[13] G. Artner *et al.*, "Measuring the Impact of Outdated Channel State Information in Interference Alignment Techniques," *Sensor Array and Multichannel Signal Process. Wksp. Proc.*, 22–25 June 2014, pp. 353–56.

[14] J. Rodríguez-Piñeiro *et al.*, "Emulating Extreme Velocities of Mobile LTE Receivers in the Downlink," *Proc. EURASIP J. Wireless Commun. and Networking*, vol. 2015, issue 106, Apr. 2015.

[15] Y. Tian and A. Yener, "Guiding Blind Transmitters: Degrees of Freedom Optimal Interference Alignment Using Relays," *IEEE Trans. Info. Theory*, vol. 59, no. 8, Aug. 2013, pp. 4819–32.

Biographies

Cenk M. Yetis [S'00, M'10] (cenkmyetis@ieee.org) received his B.Sc.'01 from Isik University, and his M.Sc.'04 and Ph.D.'10 from Istanbul Technical University, Turkey. From 2003 to 2007, he worked as an engineer at a wireless service provider in Turkey. From 2007 to 2010, he was a visiting researcher in the United States. From 2010 to 2016, he held academic positions at universities in Hong Kong, Singapore, and Turkey. In 2016, he joined Academia Sinica, Taiwan. His research interests include signal processing, information, communication, and optimization theories for wireless communications.

Jacobo Fanjul [S'13] (fanjulj@unican.es) received his telecommunication engineering (M.Sc.) degree from the University of Cantabria, Santander, Spain, in 2014. In 2013, he joined the Department of Communications Engineering, University of Cantabria, where he is currently pursuing his Ph.D. in electrical engineering. During 2016, he was a visiting researcher at the Department of Electrical Engineering and Computer Science, University of California, Irvine. His current research interests include signal processing algorithms for interference alignment, heterogeneous networks (HetNets), MIMO testbeds, and interference management for noncoherent wireless communication.

José A. García-Naya [S'07, M'11] (jagarcia@udc.es) studied computer engineering at the University of A Coruña (UDC), Spain, where he obtained his M.Sc. degree in 2005 and his Ph.D. degree in 2010. Since 2005 he is with the Group of Electronics Technology and Communications (GTEC) at UDC, where he is currently an associate professor. His research interests are in the field of wireless communication systems, with special emphasis on their experimental evaluation.

Nima N. Moghadam [S'12] (nimanm@kth.se) received his B.S. degree in electrical engineering from Shahid Beheshti University, Tehran, Iran, in 2008 and his M.S. degree in wireless systems from KTH Royal Institute of Technology, Stockholm, Sweden, in 2010, where he is currently pursuing a Ph.D. degree. His research interests lie in the area of wireless communication with emphasis on multiantenna cellular communications, radio resource allocation, and software-defined radio.

Hamed Farhadi [S'06, M'15] (farhadi@seas.harvard.edu) received his Ph.D. degree in telecommunication from KTH Royal Institute of Technology in 2014. He has been a postdoctoral research fellow at the John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, since 2016, and a researcher in the Department of Signals and Systems, Chalmers University of Technology, Gothenburg, Sweden, since 2015. He is an Associate Editor of the Springer *International Journal of Wireless Information Networks*.

Practical LTE and WiFi Coexistence Techniques Beyond LBT

Jonathan Ling, David Lopez-Perez, and Mohammad R. Khawer

ABSTRACT

Coexistence with WiFi is the key issue for unlicensed band LTE. The main coexistence mechanism is listen-before-talk, whereby radio frequency energy is sensed over a short period and compared to a threshold. Given the default energy thresholds, the energy sensing range is much less than the cell range. Either technology can experience a collision due to transmissions being below energy detection threshold. Currently, WiFi is agnostic to unlicensed band LTE. To improve coexistence, a relay-based communications channel is proposed whereby LTE announces its presence. Legacy WiFi APs may be programmed to interpret and respond by firmware upgrade . Higher performance for both networks is demonstrated via more effective radio frequency channel selection and adaptive energy detection thresholding.

INTRODUCTION

Third Generation Partnership Project (3GPP) Long Term Evolution (LTE) networks carry a huge amount of data, driven by the growing number of LTE subscribers, which reached 1 billon by mid-2016. Moreover, radio capabilities have been rapidly evolving with the development of LTE-Advanced, which has made available peak data rates of 450 Mb/s with carrier aggregation as of today.

In a new effort to accommodate the exponential growth of data traffic and further enhance user experience, the mobile industry has begun to look at unlicensed spectrum as a viable solution to improve the capacity of their networks. The 3GPP has started a new activity, known as Licensed Assisted Access (LAA) [1], to allow the usage of unlicensed spectrum alongside LTE licensed spectrum. In addition, a new industry standard, MulteFire [2], has been created to allow standalone operation of LTE-like technology in the unlicensed band, without the need for paired licensed spectrum.

LAA allows traditional operators to benefit from the additional capacity available, particularly at hotspots and in corporate environments, and complement LTE licensed operation to provide higher quality of experience. MulteFire, in contrast, allows new parties, such as verticals, to deploy and operate mobile networks without the need for expensive spectrum licenses.

An important challenge for both LAA and MulteFire deployments in unlicensed bands is coexistence with other LTE-like networks and other technologies such as WiFi [3]. When operating uLTE, short for LAA and MulteFire in this article, together with WiFi on the same band, both technologies should smoothly coexist, and the deployment of a new uLTE node should not affect the performance of existing WiFi nodes more than the deployment of a new WiFi node. However, this fair coexistence between wireless technologies for most locations and times is challenging, as different technologies have different characteristics and implement different coexistence features [1–5].

To coordinate inter-radio access technology (RAT) spectrum access in a distributed and effective manner, intelligent channel selection is necessary, such that neighboring nodes, regardless of their technology, do not reuse the same channel. WiFi channel selection schemes typically select the channel that has the least load or suffers from the least interference, relying on mining neighboring WiFi node packet headers and beacons to derive such conditions. However, WiFi nodes cannot decode uLTE messages and vice versa, as they are based on a different physical layer. This disability results in a lack of presence awareness. Each technology is not aware of the presence of the other, which hampers efficient channel assignment.

In the case where uLTE and WiFi nodes reuse the same channel, as a choice or unintentionally, their nodes can still rely on listen-before-talk (LBT) to ensure that the selected channel is shared in the time domain, as LBT is mandatory in many countries' unlicensed band regulations. When using LBT, a transmitter with data to transmit must first detect the energy across the intended transmission band. This energy detection (ED) mechanism allows the transmitter to become aware of ongoing transmissions by other nodes, and dictates whether it can access the channel or not, as a function of the detected energy with respect to a given ED threshold. However, although of low complexity, LBT does not work well in all circumstances, for example, when the information is meant to be received at background noise level, or when the nodes are distant and the received signals are below the ED threshold, which can be different for different technologies. In all these circumstances, a node wishing to transmit may sense the channel as unoccupied and interfere with another node, suddenly decreasing its signal quality and affecting its transmission. The ED

Currently, WiFi is agnostic to unlicensed band LTE. To improve coexistence, the authors propose a relay-based communications channel whereby LTE announces its presence. Legacy WiFi APs may be programmed to interpret and respond by firmware upgrade . Higher performance for both networks is demonstrated via more effective radio frequency channel selection and adaptive energy detection thresholding.

Jonathan Ling and David Lopez-Perez are with Nokia Bell Labs; Mohammad R. Khawer is with Nokia Mobile Networks CTO.

Digital Object Identifier:
10.1109/MCOM.2017.1601194

0163-6804/17/$25.00 © 2017 IEEE

Based on the neighboring cell identification, a base station, that is, an LAA enhanced NodeB or WiFi access point, may first select the "cleanest" channel, and second adjust the politeness of its MAC protocols. Self-identification is also necessary to further improve self-organizing-network capabilities.

threshold could be lowered to mitigate this issue, but arbitrarily lowering the ED threshold may significantly degrade the overall network performance by preventing simultaneous transmissions that would otherwise be successful. Moreover, a very low ED threshold may result in many false detections due to noise.

As can be derived from the above discussion, there is a need for additional mechanisms for effective inter-RAT media access control (MAC), in both the frequency and time domains. In the dynamic spectrum sensing community, explicit inter-RAT signaling is used to coordinate channel access. In some solutions, secondary clients of the band can use a channel after a database registration and lookup. However, such an approach is difficult indoors due to the limited availability of accurate position. Another approach is to design from the scratch a new common MAC protocol for both uLTE and WiFi, but that would ignore the huge installed base of 802.11 stations. Other solutions have been proposed around LTE-Unlicensed (LTE-U), the predecessor of LAA, which have the uLTE radio working in conjunction with a WiFi radio, for example, to adapt the uLTE transmission period based on WiFi traffic sensing [6] and WiFi header decoding [7]. Both schemes have demonstrated benefits, but in both cases the neighboring WiFi nodes are not explicitly aware of uLTE. As a result, the coordination is unilateral. uLTE can adapt to WiFi, but the opposite is not true, which makes coexistence suboptimal.

In this article, we propose to enhance uLTE and WiFi coexistence through a new framework that encompasses three novel mechanisms. First, we propose a new signaling framework that allows each technology to be aware of the presence of the other. Second, we propose to use such information to allow enhanced channel selection, where WiFi nodes can account for uLTE nodes. Third, when uLTE and WiFi nodes use the same channel, we propose adaptive ED threshold tuning to address the below ED threshold coordination issue.

The remainder of this article is organized as follows. The following section discusses the limitations of current signaling mechanisms and proposes two relaying techniques. Then we describe how channel selection is improved using this relaying channel, and depict a new channel selection mechanism. Following that, we discuss some limitations of LBT, and describe the problem of coexistence below ED threshold in detail. Moreover, motivated by this discussion, we then show how the LBT can be enhanced and the ED thresholds tuned to improve fairness. Finally, the "Lessons Learned" section summarizes the work, and provides thoughts on future directions.

Signaling Framework to Enhance Inter-RAT Awareness

Basic coordination of spectrum resources requires that each technology identify itself, for example, type of physical and MAC layer as well as other network features. Based on the neighboring cell identification, a base station, that is, an LAA enhanced NodeB (eNB) or WiFi access point (AP), may first select the "cleanest" channel, and then adjust the politeness of its MAC protocols. Self-identification is also necessary to further improve the self-organizing-network (SON) capabilities.

Cell discovery mechanisms in LTE and WiFi and their issues are discussed next, followed by a discussion on how to improve inter-RAT awareness via a new signaling framework

Existing Intra-Technology Cell Identification

Cell discovery in LTE is based on the physical broadcast channel (PBCH). Since user equipment (UE) attachment is network directed, the PBCH contains only the necessary information to build a connection. The master information block (MIB) contains the system bandwidth and the system frame number, and is repeated every 40 ms. The MIB is detected via autocorrelation with the primary synchronization sequence (PSS). The system information blocks (SIBs) contain additional information, which is carried on the physical downlink shared channel (PDSCH) and time multiplexed over the 40 ms slots. SIB1 contains the operator identifier (public land mobile network, PLMN) and the cell identifier, among others.

There are two types of LTE access on unlicensed frequencies: LAA, which acts as a supplemental downlink (LAA Release 13) and/or a supplemental uplink (eLAA in Release 14) to a licensed LTE carrier, and MulteFire, which is characterized by fully standalone operation in the unlicensed band. In LAA, both the licensed and unlicensed bands are operational at the same time, that is, data may be received over both bands simultaneously, but the PBCH is carried only on the licensed carrier. LTE Release 12 discovery reference signals (DRSs), which include the PSS, are transmitted at 40 ms intervals on the unlicensed carrier for time and frequency synchronization purposes. However, detection of DRS alone does not provide any information (i.e., cell ID), and one cannot even determine the operator. The licensed carrier is needed. MulteFire transmissions instead include the PBCH/PDSCH in their downlink unlicensed transmissions, now called ePBCH, which doubles the energy in the PSS and secondary synchronization signal (SSS) sequences, improving detectability.

Cell discovery in WiFi is based on the WiFi beacon, which is a message broadcast to all WiFi stations (STAs), describing various characteristics of the WiFi AP, such as its service set identifier (SSID), frequency channel, and timestamp, as well as the features it supports. WiFi beacons are repeated regularly and encoded at the lowest modulation and coding scheme (MCS). They access the channel as a priority frame, per the regular distributed coordination function (DCF) procedure.

In terms of inter-cell coordination, it is important to note that WiFi, through the 802.11aa specification, shares additional information, such as cell loading per traffic characteristic, to support improved coexistence in the so-called overlapped base station system (BSS) scenarios. However, uLTE does not have such ability yet, although an LTE beacon that provides cell loading has been proposed [8].

Proposed Inter-Technology Cell Identification

For inter-RAT cell identification, every uLTE eNB could include a WiFi receiver to decode WiFi beacons. Likewise, every WiFi AP could include a uLTE receiver to decode PBCH and PDSCH. Clearly this imposes additional costs and requirements. A better solution would be an approach that utilizes those uLTE and WiFi receivers that are naturally co-located (e.g., at UEs). Moreover, such a distributed solution should be compatible with legacy WiFi APs, in the sense of requiring at the most a software upgrade.

This distributed UE-based cell discovery approach may be taken when the eNB or AP has a smartphone with both WiFi and uLTE connected to it, as shown in Fig. 1a. This permits an apparently straightforward solution, whereby the LTE modem asks for a channel scan from the WiFi modem, and vice versa. This UE assisted scheme works best with enterprise networks that utilize such measurements in their SON algorithms. One disadvantage of the UE-based approach is that a multi-RAT UE must be available, and for optimal performance, regular scans must be taken, reducing its battery life. Moreover, multiple UEs may be necessary to detect all the neighboring cells.

To solve the mentioned UE-based approach issues, we further propose a network-based approach, in which a uLTE eNB directly communicates to a "friendly" WiFi AP. This friendly WiFi AP may be utilized, for example, to relay LTE system information and loading to surrounding WiFi networks via the WiFi beacon mechanism. We refer to this WiFi beacon carrying uLTE information as a pseudo beacon for uLTE small cells, as shown in Fig. 1b. Note that the uLTE eNB and the "friendly" WiFi AP should be at a reasonable distance, such that the pseudo beacons mostly reach the nodes that are within the uLTE eNB coverage area. Similarly, the concept of a "friendly" uLTE eNB can be used.

Enhanced Channel Selection

Channel selection algorithms, whether centralized or distributed, should provide the highest performance to end users in the long term. Alternatively, this goal may be expressed as choosing the unlicensed channel with the least activity and the least interference. Channel selection is performed infrequently, as scanning prevents traffic from being served. Some nodes may also incur service interruption if they are unable to interpret channel switch announcements.

According to the proposed network-based signaling framework, WiFi can consider uLTE while performing channel selection. As indicated earlier, the helper WiFi AP would transmit a WiFi beacon with the ID of the uLTE cell and other critical information for current channel selection algorithms such as cell loading and other information shown in Table 1. This pseudo beacon makes the uLTE cell appear as another WiFi AP to unmodified WiFi APs and STAs, providing partial backward compatibility. Note that the full interpretation of the pseudo beacon requires a firmware upgrade at the WiFi AP.

The information of Table 1 can be provided to channel selection and adaptive ED thresholding algorithms to improve the overall network coexistence and performance. Algorithm 1 further illustrates this, where U is a list of utilizations of S_f and $N_{Attached}$ is the number of attached clients of S_f.

In step 1, WiFi APs running the proposed algorithm do a channel scan, and acquire their neighboring cell list S. Step 2 filters weak nodes from S to create the filtered S_f, since the links of filtered nodes will tend to overlap and reuse the channel, despite the virtual carrier sense. Step 3 calculates the channel selection metric M, which balances current and future channel usage and dictates how the channel is time shared based on the

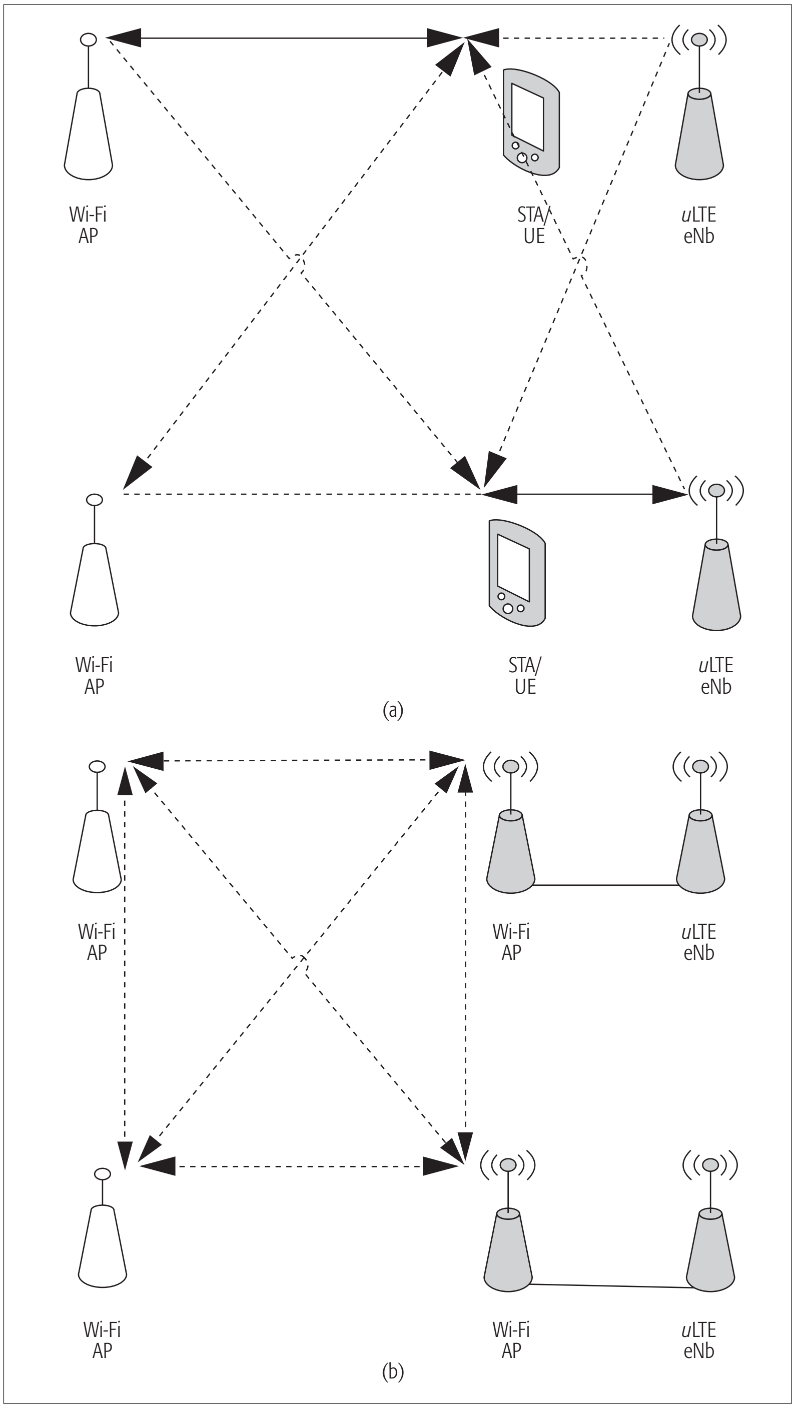


Figure 1. Inter-RAT relaying via client and via the base station: a) UE-based approach; b) network-based approach. Solid lines denote relaying channels. Dashed lines denote over-the-air detection.

uLTE parameter	WiFi beacon field	Purpose
Operator/cell ID/PLMN	SSID	Identification
Channel number	DS parameter set	Identification + channel selection
Channel number	HT operation	Identification
Loading: station count, channel utilization, available admission capability	BSS load	Load balancing/ admission control
Node type: Rel-X LAA/MulteFire/LTE-U	Vendor-specific field	Identification + channel delection
MAC spec: LBT Cat-X, other	Vendor-specific field	Identification
TX power offset (relative to AP beacon) in dBm	Vendor-specific field	Channel selection

Table 1. Example of 802.11 beacon fields populated with uLTE data.

```
1. Let $S$ be result from a scan of AP/eNBs
2. Let $S_f$ be the filtered $S$ to remove weak base stations according to their RSSI being less than threshold.
   • For uLTE eNBs, adjust RSSI according to the "TX power offset" field of Table 1.
3. Compute the channel usage based on the combined metric of actual and predicted airtime usage:
   $M_c = w_1 * \text{average}(U) + w_2 * \text{sum}(N_{\text{Attached}})$
   • For uLTE eNBs, according to the "MAC Spec" field of Table 1, adjust upward the partial metric.
4. Select the channel with the minimum metric:
   $C^* = \text{argmin}(M)$.
```

Algorithm 1 . Enhanced Channel Selection Algorithm (at AP).

weights w_1 and w_2. Finally, step 4 computes the actual channel section by solving an optimization problem that attempts to minimize the devised cost function in a greedy fashion.

It is important to note that the proposed channel selection algorithm is not a traditional WiFi one. Since uLTE and WiFi nodes may transmit at different powers, the estimated received signal strength indicator (RSSI) by a neighboring WiFi node when receiving the pseudo beacon transmitted by the helper WiFi AP may not necessarily be the same as it would have been if transmitted by the *u*LTE base station. A correction factor should be applied to such estimated RSSI according to the "TX power offset" field of Table 1. Moreover, when running on an eNB, the proposed channel selection algorithm would adjust upward the metric in step 3, causing the eNB to preferentially contend with other eNBs.

Based on the above proposed channel selection algorithm, we observe how the usage of a "friendly" WiFi AP enables the channel selection algorithm to correctly detect the base stations on each channel, whether uLTE or WiFi; and devise a metric and thereafter a channel assignment based on their reported and predicted utilization.

Limitation of Existing Coordination Techniques

The performance of wireless networks in general is determined by the radio propagation channel, which in turn is characterized by both small- and large-scale fading. Small-scale fading on the order of wavelengths exists due to multipath, whereas large-scale fading, on the order of tens of wavelengths and greater, describes spherical spreading, scattering, and absorption.

In this section, the MAC coordination mechanisms of uLTE and WiFi are compared, and their inter-RAT limitations are discussed. We then discuss how small-scale fading and large-scale fading diminish the effectiveness of LBT.

Unfairness Due to Different MAC Mechanisms

Both uLTE and WiFi MAC protocols follow regulations and implement LBT. Indeed, uLTE has adopted a CAT4 LBT with an exponential backoff mechanism, like that in WiFi, which further facilitates coexistence [9]. However, although WiFi and uLTE LBTs are very similar, some differences still remain:

- WiFi uses an ED threshold of –62 dBm, while uLTE uses an ED threshold of –72 dBm.
- WiFi augments LBT with a virtual carrier sense (VCS) mechanism, operating down to a minimum of –82 dBm (–87 dBm typical), while uLTE does not.

When using VCS, WiFi packet headers indicate for how long a transmitting node will be using the channel. These packet headers, which are encoded using the most robust MCS [6], are decoded and used by the neighboring nodes to update their network allocation vector (NAV), that is, a timeline indicating at each node when the channel is free for transmission or occupied. For example, the request-to-send/clear-to-send (RTS/CTS) mechanism reserves the channel by causing the NAV to be updated by all nodes that receive the RTS around the transmitter and the CTS around the receiver.

Due to the differences in ED threshold and VCS, WiFi will not back off to uLTE below –62 dBm, but it will back off to other WiFi transmissions up to –82 dBm or lower through the VCS mechanism. In contrast, uLTE will not back off to any technology below –72 dBm, as it only relies on ED and does not implement VCS.

Moreover, effective LBT operation requires that all transmissions be above the ED threshold, but the 802.11 DATA and ACK frames are usually transmitted at different power levels, WiFi APs at 24 dBm and WiFi STA at 14 dBm, and received at different RSSI by a uLTE eNB due to different positions. Assume the eNB follows the Cat 4 specification in 3GPP TS 36.213 and easily detects a DATA frame and refrains from transmitting, as shown in Fig. 2a. After the end of the DATA frame, the energy drops, the channel is clear, and the eNB starts a timer for 1 SIFS + 1 slot. If the ACK is detected, that is, if it is received above –72 dBm, the eNB will refrain from transmitting. If the ACK is received below –72 dB, the eNB can transmit, as shown in Fig. 2b. Therefore, collisions are possible whereby an ACK is in the process of being received, while the eNB does not detect it and goes on to transmit, as shown in Fig. 2c.

The ED threshold of –72 dB effectively limits the collision-free downlink range of the neighboring AP, that is, the WiFi STA must be close enough to both the WiFi AP and uLTE eNB such that the ACK is received at –72 dBm or greater. Indeed, this mid- range interference zone at below ED threshold, from –72 dBm to –87 dBm, represents a grey area for coexistence.

Effect of Small-Scale Radio Propagation on LBT

Coordination through ED requires all nodes to receive all transmitted signals above the ED threshold all the time, that is,

$$P\{\mathit{ED\ success}\} = \prod P\{P_t G(n, m) > \mathit{ED\ threshold}\}$$

where P_t is the transmit power, and $G(n, m)$ is the path gain from node n to node m, considering both small- and large-scale fading.

This is a difficult requirement to satisfy. Let us assume there are 5 nodes wanting to access the channel, that is, there are 10 links, counting reciprocal links as a single link, and that the small-scale fading power is Chi-square distributed. Notably 10 percent of the time there is a 10 dB or greater fade. Given the above fading channel model, and assuming an ED threshold of –62 dBm, and that all nodes receive each other at –52 dB on average, $P\{\mathit{ED\ success}\} = (.90)^{10}$ = 34 percent. This shows that sensing errors are prevalent due to small-scale fading even for a small number of nodes and when the average signal strength is much higher than the ED threshold.

Effect of Large-Scale Radio Propagation on LBT

For the default ED thresholds, sensing failure of the other technology's transmission may be expressed as

P{*ED failure at uLTE*} =
P{RSSI < –72 dBm at eNB | STA transmits} and
P{*ED failure at WiFi*} =
P{RSSI < –62 dBm at AP | UE transmits}.

To quantify these probabilities let us assume a single eNB and AP, closely located to each other, each with a single client. Let clients be uniformly located in a simulated building of 50 × 120 m. The bases are located off center at (25,30) m. Large-scale propagation for an "open" building is modeled by the InH propagation model in 3GPP TR 36.814, while large-scale propagation for a "light partitioned" building is modeled by a diffusion model with its parameters in [10].

Figure 3 provides the cumulative density function of the RSSI for both propagation models, assuming 14 dBm transmit power and 6 dB total antenna gain. Vertical lines show the minimum WiFi signal strength of –87 dBm [3] and the minimum LTE signal strength of –100 dBm [11]. ED thresholds are also highlighted at –62 dBm and –72 dBm for WiFi and uLTE, respectively. Applying the minimum signal strength threshold, WiFi cell coverage is 87 percent by InH and 62 percent by diffusion. Likewise, uLTE coverage is 100 percent by InH and 75 percent by diffusion, which reveals the impact of large-scale fading.

The sensing failure probability can be obtained from the CDF in Fig. 3 by renormalizing it to the coverage area, so P{*ED failure at WiFi*} = P{RSSI <= –62 | UE RSSI>–100} = 1 – P{ RSSI > –62}/P{RSSI > –100} = 56 percent for InH, and 73 percent for the diffusion model. P{*ED failure at uLTE*}= P{RSSI <= –72 | RSSI > –87} = 42 percent for InH and 33 percent for the diffusion model. This shows that there is a large area where signals are received below ED threshold, and inter-RAT coordination is not possible relying on ED alone.

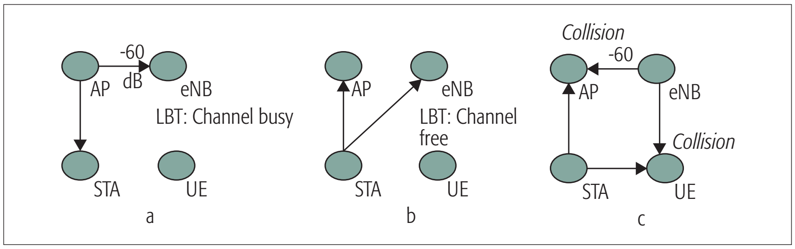


Figure 2. Collision scenario: a) downlink transmission received above ED; b) uplink ACK received below ED; c) collision at AP and UE due to transmitting eNB.

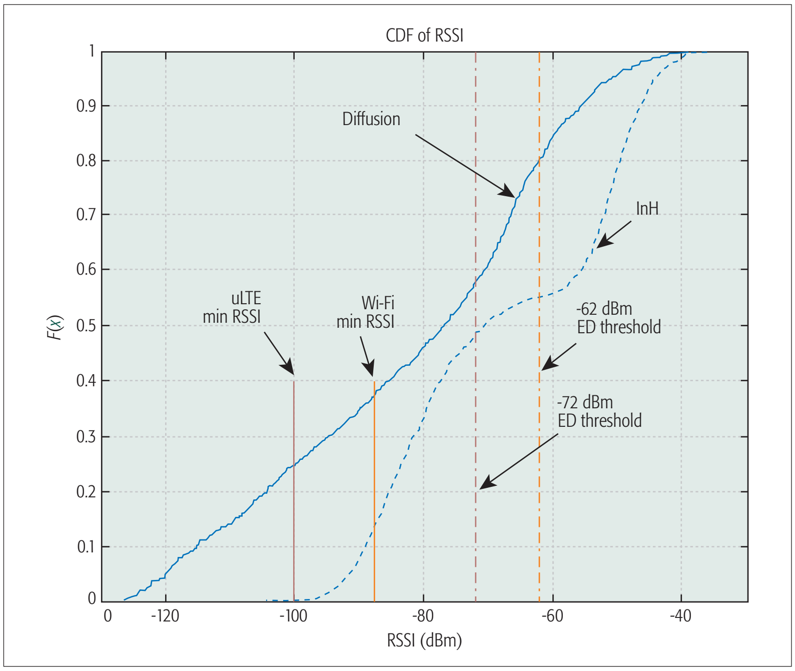


Figure 3. CDF of RSSI for two indoor propagation models and TX power of +14 dBm (6 dB antenna gain).

Addressing the Below ED Challenge

Following the channel selection process, we propose that the ED thresholds are adapted to improve efficiency and fairness when both uLTE and WiFi share the same channel.

On the uLTE side, Algorithm 2 works in a periodical manner (note that a similar algorithm can be applied on the WiFi side with the appropriate thresholds):

Note that the values of $\mathrm{RSSI}_w^{\mathrm{WiFi}}$ and $\mathrm{RSSI}_j^{\mathrm{uLTE}}$ are averaged using a filter to mitigate fast-fading effects and smooth the measurements.

We simulated the network performance of an LAA cell coexisting with a WiFi cell in the same unlicensed channel of 20 MHZ in the 5 GHz band using the proposed adaptive ED thresholding algorithm. The performance evaluation is conducted over an enterprise scenario of 120 m × 50 m, where there is a uLTE eNB located at position (30,25) m and a WiFi AP located at position (90,25) m, 60 m apart from each other. It is important to note that the InH channel model is used in this case, but that the link between the LAA eNB and the WiFi AP is always set to be non-line of sight and below the default ED detection

1) The ED threshold Γ of the eNB of interest is initially set to –72 dBm, as per today's operation where *i* denotes the eNB of interest. A minimum ED threshold Γ^{min} is also set, for example, –82 dBm.
2) The eNB periodically determines the RSSI, $RSSI_w^{WiFi}$, of the received beacon signal from each of the co-existing WiFi nodes *w*. Then, the eNB finds the minimum one, RSS_{min}^{WiFi}.
3) During idle periods, the eNB also looks for other eNBs using MIB/SIB/CRS decoding. If this capability is present, the eNB periodically determines the RSSI, $RSSI_j^{uLTE}$, of the received MIB/SIB/CRS from each of the co-existing uLTE j. Then, the eNB finds the minimum one, RSS_{min}^{uLTE}.
4) Thereafter, the eNB determines the minimum received RSSI, that is, $RSSI_{min} = \min\{RSS_{min}^{WiFi}, RSS_{min}^{uLTE}\}$.
5) Then, the ED threshold of eNB is set as follows:

$$\Gamma = \begin{cases} -72\text{ dbm}, & RSSI_{min} \geq -72\text{ dbm} \\ RSSI_{min}, & \Gamma^{min} < -72\text{ dbm} \\ G^{min}, & RSSI_{min} \leq \Gamma^{min} \end{cases}$$

6) If for a period of time T, the eNB does not detect the presence of any other technology nodes on the unlicensed channel, or their RSSI value is above –72 dBm, it will set the ED threshold back to –72 dBm.

Algorithm 2. Adaptive ED Thresholding Algorithm.

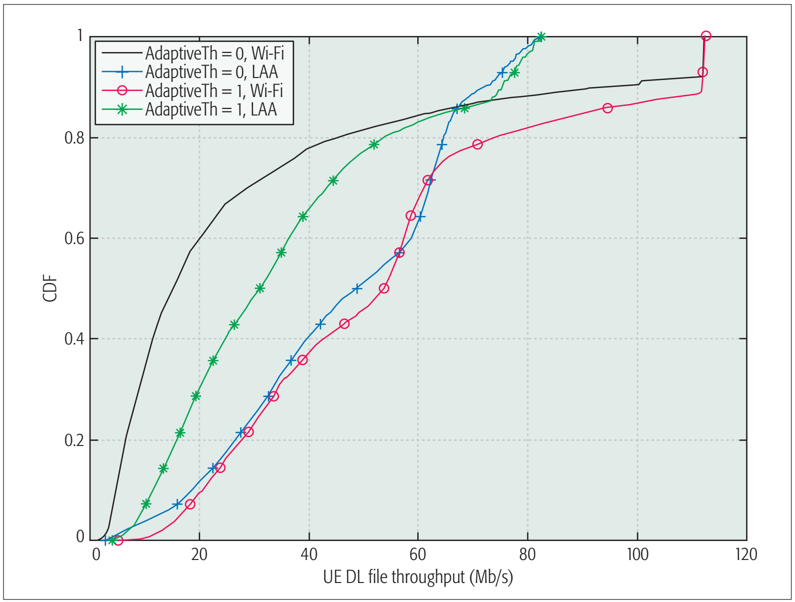


Figure 4. Downlink file throughputs.

threshold. One LTE UE and one WiFi STA located at the cell edge of their respective servers are considered. The UE and STA use a downlink FTP service, following the 3GPP FTP traffic model 3, where the FTP file size is 2 MB and the packet arrival rate is Poison distributed with an average of 0.625. 2 × 2 multiple-input multiple-output (MIMO) is considered, and 64 quadrature amplitude modulation (QAM) is the maximum modulation scheme supported. The transmission opportunity is set to 3 ms for both technologies. Note that LAA can only start transmission at the subframe boundary, and a reservation signal is used from the moment LAA acquires the channel to such subframe boundary to guarantee that there are no collisions with WiFi. It is important to note that due to the nature of FTP traffic and because it has been shown to improve LAA and WiFi coexistence, the RTS/CTS mechanism is enabled in our simulations. 100 simulation drops are performed, and in each drop 10 s are simulated. Please refer to [12] for a more complete description of the simulator.

Figure 4 shows the results in terms of UE/STA downlink file throughput. Without the adaptive ED scheme, the *u*LTE eNB and the WiFi AP do not detect each other, and thus their downlink transmissions are not coordinated. This results in a high number of collisions and re-transmissions, which mostly affect WiFi performance. This is because WiFi quickly detects the collision through the RTS/CTS mechanism and continuously backs off, while uLTE, which does not have RTS/CTS, continues to transmit, temporarily forcing WiFi off of the band. Eventually, the WiFi AP transmits to its STA, when the uLTE eNB does not have any data to transmit to its UE. In contrast, with the adaptive ED scheme, the WiFi AP and the uLTE eNB sense each other's transmissions, and are thus able to coordinate. This results in fairer time sharing, as well as much fewer collisions and retransmissions. Such coordination benefits WiFi as it does not back off as much to uLTE, but necessarily impacts uLTE as it decreases its air time to share with WiFi.

In terms of median throughput, WiFi performance increases 3.5× (from 15 Mb/s to 54 Mb/s), while uLTE decreases 36 percent (from 49 Mb/s to 31 Mb/s). It is important to note that the overall system performance increases with the proposed adaptive ED threshold scheme by 32 percent (from 64 Mb/s to 85 Mb/s). In short, with adaptive thresholding, the channel is shared more fairly, and the overall efficiency has improved.

Lessons Learned

Simple LBT provides limited inter-RAT coordination, as below ED threshold signals occur in a large fraction of the cell area. A relaying channel based on co-located modems is proposed. With network-based relaying, WiFi pseudo beacons make uLTE cells appear as another WiFi AP to unmodified APs. This provides partial backward compatibility, and the full interpretation of beacon information only requires a firmware upgrade at the AP. This facilitates channel selection and adapting MAC parameters as both technologies may now positively identify each other's presence. Simulations show improvement in throughput when uLTE and WiFi adapt their ED thresholds and coordinate with each other.

As usage of both LTE and WiFi technologies continues to grow, coordination beyond LBT will become necessary to maintain fairness and provide high quality of experience. A local inter-RAT channel will play a strong role in the effectiveness of MAC protocols and coordination.

As usage of both LTE and WiFi technologies continues to grow, coordination beyond LBT will become necessary to maintain fairness and provide high quality of experience. A local inter-RAT channel will play a strong role in the effectiveness of MAC protocols and coordination.

References

[1] A. Mukherjee *et al.*, "Licensed-Assisted Access LTE: Coexistence with IEEE 802.11 and the Evolution toward 5G," *IEEE Commun. Mag.*, vol. 54, no. 6, June 2016, pp. 50–57.

[2] Multefire; http://www.multefire.org/, accessed May 15, 2017.

[3] E. Perahia and R. Stacy, *Next Generation Wireless LANs*, Cambridge Univ. Press, 2nd ed., 2013.

[4] A. Cavalcante *et al.*, "Performance Evaluation of LTE and WiFi Coexistence in Unlicensed Bands," *Proc. IEEE VTC-Spring*, June 2013, pp. 1–6.

[5] F. M. Abinader *et al.*, "Enabling the Coexistence of LTE and WiFi in Unlicensed Bands," *IEEE Commun. Mag.*, vol. 52, no. 11, Nov. 2014, pp. 54–61.

[6] Qualcomm Technologies Inc, "LTE-U Technology and Coexistence," May 28, 2015; http://www.lteuforum.org/uploads/3/5/6/8/3568127/lte-u_coexistence_mechansim_qualcomm_may_28_2015.pdf, accessed May 15, 2017.

[7] A. Bhorkar *et al.*, "Medium Access Design for LTE in Unlicensed Band," *Proc. IEEE Wireless Commun. Networking Conf. Wksps.*, New Orleans, LA, 2015, pp. 369–73.

[8] Kyocera, R1-151055, "Inter-operator LAA Cells Coordination using the LTE Beacon and the LTE Header Channels," 3GPP TSG RAN WG1 Ad Hoc Meeting, Paris, France, 24–26 March 2015.

[9] 3GPP TS 36.899, "Study on Licensed-Assisted Access to Unlicensed Spectrum," Rel. 13, May 2015.

[10] D. Chizhik, J. Ling, and R. A. Valenzuela, "Radio Wave Diffusion Indoors and Throughput Scaling with Cell Density," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, Sept. 2012, pp. 3284–91.

[11] S. Sesia, I. Toufik, and M. Baker, *LTE: The UMTS Long Term Evolution from Theory to Practice*, Wiley, 2nd ed., 2011.

[12] D. López-Prez *et al.*, "Boosted WiFi through LTE Small Cells: The Solution for an All-Wireless Enterprise," *Proc. 2016 IEEE PIMRC*, Valencia, Spain, 2016, pp. 1–6.

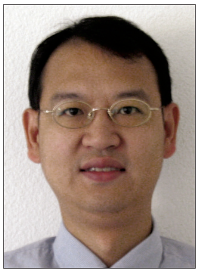
Biographies

JONATHAN LING is a researcher in Nokia Bell Laboratories currently focusing on multi-radio connectivity for 5G. His experience includes developing enhancements, evaluating performance of wireless systems, and over 40 publications. The WiFi Boost invention he demonstrated at MWC 2015 led to 3GPP LWIP. He earned a B.Sc. degree in electrical engineering from RutgersUniversity, Piscataway, New Jersey, and an M.Sc. in computer science and a Ph.D. in electrical engineering, both from Stevens Institute of Technology, Hoboken, New Jersey.

MOHAMMAD R. KHAWER is a director in 5G Leadership, and a member of the chief architect group in Nokia's Mobile Networks CTO Organization. He is responsible for driving forward disruptive technology innovations by transforming applied research ideas into viable commercial products. In April 2017, he received the prestigious Nokia Innovation Award for one such endeavor. He received his M.S. and Ph.D. degrees in computer science engineering from Syracuse University, New York. He currently holds 35 worldwide granted patents.

DAVID LÓPEZ-PÉREZ [M'12, SM'17] is a member of technical staff at Nokia Bell Laboratories, where he works primarily on ultra-dense networks and massive MIMO. He received his Ph.D. degree in wireless networking from the University of Bedfordshire, United Kingdom, in 2011. He has authored more than 100 book chapters and journal and conference papers, all in recognized venues, holds over 36 patents applications, received the IEEE ComSoc Best Young Professional Industry Award in 2016, and is an Editor of *IEEE Transactions on Wireless Communications*.

Integrated Circuits for Communications



Charles Chien

Zhiwei Xu

In this issue of the Integrated Circuits for Communications Series, we have selected two papers that highlight recent progress in the integrated circuits design of 3D imagers based on optical and microwave circuits to support high resolution imaging for various applications such as in-home spectroscopy for personal health diagnostics to meet the needs of aging populations. Interestingly, these imagers can be viewed as self-contained communication systems that transmit a sounding signal and receive the return to estimate the physical dimensions of nearby objects.

Since the emergence of 3D imaging in the 1990s, it has been adopted in many applications, such as medical diagnostics and topographic surveys. The demand for accurate 3D imaging has escalated in recent years due to industrial focus on self-driving cars, drone-based delivery, and point-of-care (PoC) systems. One method of imaging is based on light detection and ranging (LIDAR), which employs light in the form of pulsed or continuously modulated lasers to achieve precise ranging. Compared to ultrasonic and radar techniques, it leverages the short wavelength in the electromagnetic spectrum to achieve much higher resolution.

LIDAR-based 3D imagers consist of photonic elements and electronics processors. Traditionally, bulky photonic components have prevented the form-factor reduction needed for portable devices. In the past, various materials have been tried to fabricate photonic integrated circuits, such as lithium niobate, gallium arsenide, and silicon on insulator. While these materials possess properties that enable implementation of high-performance photonic systems, they do not offer a path to the integration needed for portability.

Silicon is currently the primary material used to fabricate integrated electronics with more than billions of transistors per square millimeter. Recently, researchers have endeavored to leverage silicon technology to manufacture more integrated optical components, such as photo detectors, laser diodes, optic modulators, and optic waveguides. These efforts are paving the way for small form-factor electronic-photonic imaging systems, making integrated LIDAR imagers a reality.

In the article "LIDAR System Architecture and Circuits," the authors give an architecture overview highlighting key design challenges to implement LIDAR systems in silicon technology. The article covers four LIDAR architectures with implementation examples. These include pulsed, amplitude-modulated, continuous-wave, and frequency-modulated continuous-wave (FMCW) LIDARs. The authors give insightful comparisons on the design trade-offs among axial precision, field of view, lateral resolution, operating range, and power efficiency for each LIDAR architecture. With careful design considerations, the authors have developed a fully integrated 3D FMCW LIDAR implemented in two chips that are integrated into a small form factor. The chip set consists of a silicon-photonic chip and a 0.18 μm complementary metal oxide semiconductor (CMOS) chip, stacked using through-silicon-vias (TSVs). This integrated LIDAR imager achieves 250 μm lateral resolution and 11 μm range precision to enable high resolution 3D imaging.

In contrast to photonics, microwave runs at a lower frequency and offers an alternate method to detect and diagnosis materials through impedance spectroscopy, which has emerged as a prime technology that would enable future PoC systems. The idea of conducting health diagnosis at home and personalized medicine is no longer just a concept, but has evolved into a potential high-growth market. Fortunately, impedance spectroscopy utilizes frequencies ranging from 3 kHz to 300 GHz, well within the reach of CMOS technology. This has spurred intense effort in the research community to develop highly integrated CMOS spectrometers with excellent sensitivity while consuming low power, enabling portable medical diagnostic devices.

The second article, "Integrated Circuit Technology for Next Generation Point-of-Care Spectroscopy Applications," describes spectroscopy based on nuclear magnetic resonance (NMR) and electron spin resonance (ESR). NMR exploits the interaction between the intrinsic magnetic moment of nuclei and an external magnetic field to detect specific proteins and structural changes in proteins. ESR leverages the interaction between the spin of an electron and an externally applied magnetic field to obtain spectroscopic information used to detect diseases, such as cancer, Alzheimer's, and Parkinson's disease. Traditionally, NMR and ESR instruments are bulky and consume large amounts of power, making them unsuitable for PoC applications. In this article, the authors leverage the integration capability and massive digital processing offered by CMOS technology to implement NMR and ESR spectrometers on a chip occupying only several millimeters in area. Their research exemplifies efforts underway to realize handheld NMR or ESR devices with reduced power consumption and cost for future personalized medicine and home diagnostics.

We would like to take this opportunity to thank all the authors as well as reviewers for their contributions to this Series. Future issues will continue to cover circuit technologies that are enabling new emerging communication systems. If you are interested in submitting a paper, please send your paper title and an abstract to either of the Series Editors for consideration.

Biographies

Charles Chien (charles.chien@creonexsystems.com) is the president and CTO of CreoNex Systems, which focuses on technology development for next generation communication systems. Previously, he held key roles at Conexant Systems, SST Communications, and Rockwell. He was also an assistant adjunct professor at the University of California Los Angeles. His interests focus mainly on the design of system-on-chip solutions for communication systems. He has published in various journals and conferences, and has authored a book, *Digital Radio Systems on a Chip*.

Zhiwei Xu is currently with Zhejiang University, working on cognitive radios, high-speed ADC, and mmWave ICs. He held industry positions with SST Communictions, Conexant Systems, NXP, and HRL Laboratories, where he developed wireless LAN and SoC solutions for proprietary wireless multimedia systems, CMOS cellular transceivers, multimedia over cable systems, TV tuners, software defined radios, and analog VLSI. He has published in various journals and conferences, three book chapters, and 12 granted patents.

Lidar System Architectures and Circuits

Behnam Behroozpour, Phillip A. M. Sandborn, Ming C. Wu, and Bernhard E. Boser

Abstract

3D imaging technologies are applied in numerous areas, including self-driving cars, drones, and robots, and in advanced industrial, medical, scientific, and consumer applications. 3D imaging is usually accomplished by finding the distance to multiple points on an object or in a scene, and then creating a point cloud of those range measurements. Different methods can be used for the ranging. Some of these methods, such as stereovision, rely on processing 2D images. Other techniques estimate the distance more directly by measuring the round-trip delay of an ultrasonic or electromagnetic wave to the object. Ultrasonic waves suffer large losses in air and cannot reach distances beyond a few meters. Radars and lidars use electromagnetic waves in radio and optical spectra, respectively. The shorter wavelengths of the optical waves compared to the radio frequency waves translates into better resolution, and a more favorable choice for 3D imaging. The integration of lidars on electronic and photonic chips can lower their cost, size, and power consumption, making them affordable and accessible to all the abovementioned applications. This review article explains different lidar aspects and design choices, such as optical modulation and detection techniques, and point cloud generation by means of beam-steering or flashing an entire scene. Popular lidar architectures and circuits are presented, and the superiority of the FMCW lidar is discussed in terms of range resolution, receiver sensitivity, and compatibility with emerging technologies. At the end, an electronic-photonic integrated circuit for a micro-imaging FMCW lidar is presented as an example.

The authors explain different lidar aspects and design choices, such as optical modulation and detection techniques, and point cloud generation by means of beam-steering or flashing an entire scene. Popular lidar architectures and circuits are presented, and the superiority of the FMCW lidar is discussed in terms of range resolution, receiver sensitivity, and compatibility with emerging technologies.

Introduction

The dream of self-driving cars has finally become a reality, but not yet a commodity. In addition to legal barriers, many technical issues remain to be solved before the steering wheel can be abandoned. Among these technical challenges is the refinement of the 3D imaging and mapping tools used for object recognition, navigation, and collision avoidance. The performance offered by processing 2D images, such as stereovision techniques, would not be sufficient for these purposes, necessitating the use of direct rangefinders based on ultrasonic, radar, and lidar technologies. The propagation of ultrasonic waves through air induces large losses that prevent the waves from reaching distances beyond a few meters, whereas radar and lidar waves can both propagate across long distances. Radar is a well established tool that can work even in poor weather conditions such as heavy rain, snow, or fog. However, the shorter wavelength and superior beam properties of the lightwaves used in lidar offer a more suitable choice for 3D imaging and point cloud generation. Unfortunately, current lidar solutions are costly, bulky, and power-hungry, or they perform poorly. Researchers in this area are working to develop an inexpensive solution that offers the required performance with reasonable size and power consumption. In addition to self-driving cars, numerous other applications will benefit from an affordable 3D imaging technology: Drones that need 3D imaging for their navigation are becoming increasingly popular for use in surveillance, delivery of goods, aerial mapping, agriculture, construction, high-risk monitoring, defense, and search and rescue missions. The number of personal and industrial robots is projected to surpass tens of millions during the next decade, and 3D imaging could become a popular aid for their control. There are countless other applications in medical, scientific, industrial, defense, and consumer areas that would benefit from lidar-based rangefinders and 3D cameras.

Specifications including the operating distance, range resolution, acceptable ambient light and interferers' levels, measurement speed and frame rate, multi-target detection capability, power consumption, maximum permissible optical exposure, and other parameters can vary significantly across different applications. This article describes the basic lidar architecture, followed by more details on popular lidar schemes that provide insight into the important design choices and trade-offs.

Basic Lidar Architecture

Figure 1 illustrates the main components of a typical lidar, which, like radar, includes a transmitter and a receiver. The range R is measured based on the round-trip delay of light to the target, τ:

$$R = \frac{1}{2} c \cdot \tau \tag{1}$$

where c is the speed of light in the medium between the lidar and the target (e.g., air). Based on this equation, and because in most cases the speed of light is known to a very good accuracy, the lidar-based range measurement is equivalent to measuring the round-trip delay of lightwaves to the target. This is achieved by modulating the intensity, phase, and/or frequency of the waveform of the transmitted light and measuring the time required for that modulation pattern to appear back at the receiver. In the most trivial case of intensity modulation, a short light pulse

Behnam Behroozpour is with Bosch LLC; Phillip A. M. Sandborn, Ming C. Wu, and Bernhard E. Boser are with the University of California at Berkeley.

Digital Object Identifier: 10.1109/MCOM.2017.1700030

0163-6804/17/$25.00 © 2017 IEEE

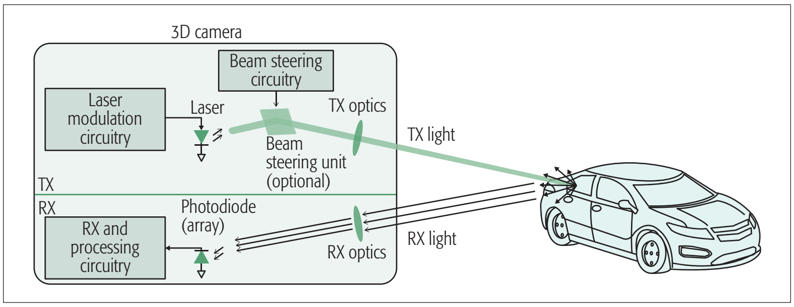


Figure 1. Basic lidar-based 3D camera architecture.

is emitted toward the target, and the arrival time of the pulse's echo at the receiver marks the distance. Lasers are the preferred source of light because of their narrow spectra and superior beam properties; furthermore, phase- and frequency-modulation (PM and FM) lidars require the laser light's coherence. Lasers with wavelengths of 905, 1300, or 1550 nm, which are near the three established telecommunications windows, are commonly used in lidar applications.

To create a 3D image, the light should be directed to all the points in a desired field of view (FOV). This can be done by distributing the light to the entire scene at once (flash lidar), by employing a beam-steering unit to scan the FOV, or by a combination of these. In flash lidar, the different points in the FOV should be differentiated in the receiver using proper imaging optics, similar to the lens-set of a photographic camera. Over the years, many different beam-steering techniques have been developed. Foremost among these are mechanical motion of the light source [1]; deflection of the light using a macro- or micro-mechanical mirror [2]; optical-phased arrays (based on liquid crystals [3], MEMS mirrors [4], or silicon-photonic tunable phase elements [5] and wavelength tuning [6]).

Finally, in the receiver, the scattered light from the target is collected, and the delay in its modulation pattern vs. the source light is extracted and used for ranging. In a 3D camera based on flash lidar, the receiver has multiple pixels, and the time of flight should be measured separately for each pixel.

Important Performance Metrics

The most important performance metrics for a lidar-based 3D camera are its axial precision, lateral resolution, FOV, frame rate, transmit power in relation to eye safety, maximum operating range, sensitivity to ambient light and interferers, power consumption, and cost. These metrics are briefly discussed here.

Axial Precision

The terms axial or range precision refer to the standard deviation of multiple range measurements performed for a target at a fixed distance (σ_R). This should not be confused with range resolution (δR). Range resolution refers to the lidar's ability to resolve multiple closely spaced targets in the axial direction. For example, when 3D imaging an organic tissue, the emitted light is reflected by the interfaces between the tissue's different layers. In this case, better axial resolution helps in detecting thinner tissue layers, while better axial precision improves the certainty with which the interfaces between these layers can be located. The latter can be improved by averaging the results of multiple measurements.

For any time-of-flight ranging system based on either electromagnetic or ultrasonic waves, the range resolution can be found using the following equation [7]:

$$\delta R = \frac{c}{2B} \tag{2}$$

where c is the velocity of the waves, and B is the bandwidth of the information they carry. This means the information content on the waves should vary fast enough that the reflections from two targets separated by δR can be meaningfully distinguished in the receiver. The time difference between the reflections from two such targets is $\delta t = 2\delta R/c$, translating to a bandwidth inversely proportional to this time, or $B = c/2\delta R$. The speed of the waves in air for both optical and radio frequency waves is equal to 3×10^8. The bandwidth of radio frequency waves can reach tens of gigahertz, resulting in centimeter-range resolution; however, optical waves can have much larger bandwidth, enabling micrometer-range resolution. Although the ranging precision is different from the resolution, their values are not entirely independent. In [7], it has been shown that $\sigma_R^2 \propto \delta R^2/SNR$; where σ_R^2 is the variance of the measured range, and SNR is the signal-to-noise ratio of the received signal. In other words, sharper changes in the information content of the waves, resulting in smaller δR, as well as higher SNR can improve the ranging precision.

FOV and Lateral Resolution

FOV is usually specified with two horizontal and vertical angles around the axis perpendicular to the camera aperture within which the distance can be measured. Lateral or angular resolution of a 3D camera is a measure of its ability to distinguish two adjacent points in the FOV. Optical waves with micrometer wavelength can achieve lateral resolutions of 0.1° with aperture sizes of only a few hundred micrometers ($\theta \propto \lambda/D_{aperature}$) that easily fits on a single chip. However, radio frequency waves with frequencies near 100 GHz would require a 1-m aperture for the same resolution, which is challenging to implement in many applications. In a flash lidar, similar to a photographic camera, the lateral resolution and FOV are defined by the optical front of the receiver and also the photodetector's array size. However, in a beam-steering lidar the properties of the emitted laser beam, such as its divergence angle, side lobes, and scan range, have more significant effect on the FOV and lateral resolution. FOV is of particular importance in 3D mapping for self-driving cars and drones, where a 360° view of the surroundings is often necessary. At the time of this writing, such a large FOV can be achieved either by mechanically moving a 3D camera with a smaller FOV or by stitching the outputs of multiple 3D cameras using computer software.

Emitted Power and Eye Safety

For lidar applications where a longer operating range is important, a larger transmit power is desired. However, the maximum transmit power is often limited by eye safety regulations. This is a greater concern for lidars than radars because a coherent laser beam with milliwatts of power can cause serious damage to the human eye. The *maximum permissible exposure* (MPE) of a laser product depends strongly on its wavelength, beam diameter and divergence, beam motion, duration of exposure for continuous-wave operations, and pulse width and repetition rate for pulsed operations. As a result, eye safety is an important determinant in the selection of such parameters when designing a lidar.

Maximum Operating Range

Maximum operating range is usually limited by the transmit power level and the receiver sensitivity. In a beam-steering lidar, the operating range can be improved by reducing the beam divergence and its side-lobes. In all lidar categories, a larger receive aperture can increase the amount of collected optical power and improve the operating range.

In long-range 3D cameras, beam-steering lidars are more commonly employed than flash lidars. This seems to be a straightforward choice considering that in a beam-steering lidar the entire laser power is focused on a single spot at one time, creating a stronger echo compared to the distributed light in a flash lidar; however, it must be noted that in a flash lidar, the parallel measurement of all pixels allows a longer measurement time per pixel to achieve the same frame rate, which can be used to average the noise and retain the SNR to some extent.

In the lidar types in which the modulation is applied to the phase or frequency of the light, phase noise of the laser beam can also limit the maximum operating range.

Popular Lidar Architectures

The combination of choices available for the different lidar blocks can result in a wide variety of lidar architectures. Among these, pulsed, amplitude-modulated continuous-wave (AMCW) and frequency-modulated continuous-wave (FMCW) are the most popular schemes, and these are discussed in this section. Figure 2 illustrates the range precision vs. maximum operating distance for lidars presented since 1990, and the regimes in which each of these lidar types have often been employed are indicated by the shaded areas.

Pulsed lidar can provide moderate precision over a wide window of ranges. This is thanks to the fact that the nanosecond pulses used in these lidars often have high instantaneous peak power that can reach far distances while maintaining low average power below the eye-safe limit. Furthermore, according to Eq. 2, the large bandwidth associated with short pulses can enable high-precision range measurements with a relative range error acceptable even at short distances.

AMCW lidar can achieve precision similar to that of the pulsed lidar but only at moderate ranges; it is usually secondary parameters such as the fabrication cost that motivate the selection of one or the other in this regime of range and precision. AMCW lidars are not popular for long-range measurements because they transmit continuous optical power that has to remain below the eye-safe limit at all times; therefore, the eco signal at their receiver coming from far objects is not as strong as it is in pulsed lidars.

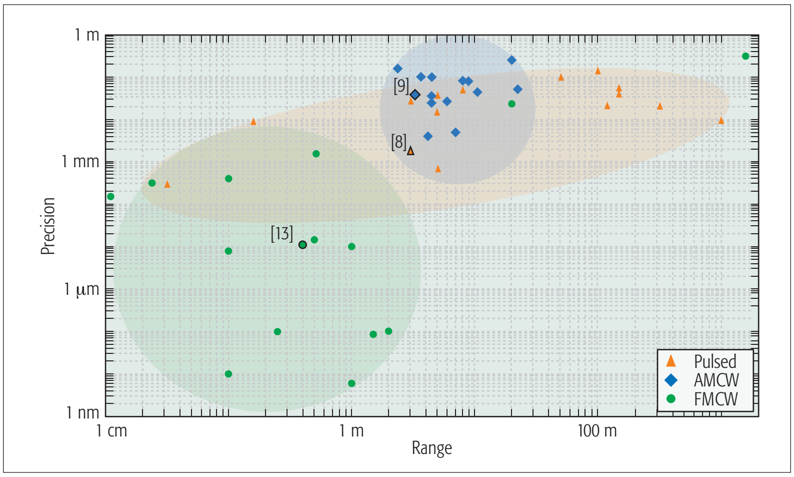


Figure 2. Precision vs. operating range for academically published and industrial lidars since 1990.

FMCW lidar is the only architecture that has been used to achieve sub-micrometer precision in multiple designs. This is enabled by direct modulation and demodulation of the signals in the optical domain with much larger bandwidth than that possible when using electronic circuitry. There are also instances of using FMCW lidar for moderate and long-range applications with a precision comparable to or better than that of pulsed and AMCW lidars.

In the rest of this article, the three popular lidar categories are discussed in more detail, and one instance in which integrated circuits were effectively used to achieve a significant performance improvement is presented for each type. These examples are highlighted in Fig. 2.

Pulsed Lidar

In this type of lidar (Fig. 3), the round-trip delay of a short pulse of light to the target is measured to find the target's distance. Shorter pulse widths are desired to increase the peak power while maintaining the average eye-safe exposure. Furthermore, from Eq. 2, it can be seen that the axial resolution of the lidar is improved by increasing the pulse bandwidth, which is equivalent to reducing its width. Most applications use pulses with durations from less than 1 ns to tens of nanoseconds.

Although the name "pulsed lidar" is mainly descriptive of the modulation method in the transmitter, it also influences the receiver design. Single-photon avalanche detectors (SPADs) are often employed in pulsed lidar receivers to improve their sensitivity and increase their operating distance. The high interest in these detectors has motivated their development in complementary metal oxide semiconductor (CMOS)-compatible processes to reduce their cost [8]. SPADs are essentially avalanche photodiodes operating in the reverse-biased mode slightly beyond their breakdown voltage. Because of the strong electric field from the reverse-bias voltage, the electron-hole pairs generated by photon absorption or thermal fluctuation are accelerated to a level that can trigger an avalanche process. At this point, the electronic circuitry around the SPAD must

Maximum operating range is usually limited by the transmit power level and the receiver sensitivity. In a beam-steering lidar, the operating range can be improved by reducing the beam divergence and its side-lobes. In all lidar categories, a larger receive aperture can increase the amount of collected optical power and improve the operating range.

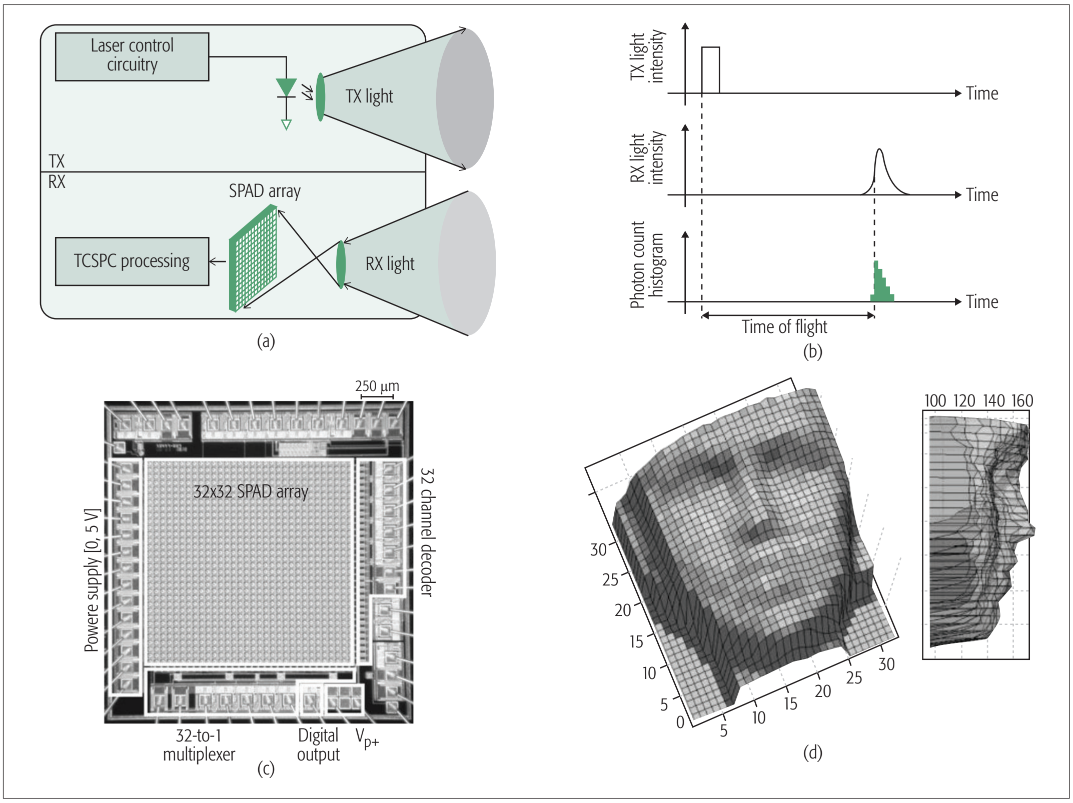


Figure 3. Pulsed lidar with flash light distribution presented in [8]: a) simplified architecture; b) timing diagram; c) chip photomicrograph; d) 3D image of a human face (in millimeters).

reduce the reverse-bias voltage to stop the avalanche and prepare the device for the next detection. The timing of the avalanche event can then be recorded by the electronic circuits to mark the arrival time of the pulse echo to the receiver. The SPAD recovery time can extend up to 100 ns and limit the measurement rate.

SPADs are susceptible to false detections due to either the thermal noise of the detector itself or photons from the ambient light that happen to be at the detectable wavelength window. Therefore, SPAD receivers are often employed in a statistical architecture where the arrival times of multiple repetitive pulses, sometimes recorded by many SPADs in parallel, are accumulated in a histogram. The recordings in a time window of comparable duration to the emitted pulse width have a higher chance of being part of the expected signal. This fact is used to filter out unwanted recordings and improve the measurement precision. This technique is referred to as time-correlated single-photon counting (TCSPC), and has gained popularity in pulsed lidars and also in fluorescence lifetime measurements.

Pulsed lidars can operate in either flash or beam-steering modes. The latter is often the preferred choice for long-range applications. Among the beam-steering technologies mentioned in previous sections, silicon-photonic phased arrays (SPPAs) are more popular because of their compatibility with fully integrated chip-scale lidars. The foremost attraction of this technology is that it could provide solid-state lidars with no mechanical parts, taking advantage of the potential high-volume and low-cost manufacturing achieved by today's integrated circuit industry. Recently, there have been preliminary demonstrations of such technologies, and strong growth in this direction is expected within the next decade. However, the large peak power of the pulsed lidars combined with the small effective cross-section of the silicon-photonic waveguides can enhance undesirable nonlinear optical processes in the silicon. Therefore, pulsed lidars are not currently preferred for use with SPPAs, compared to continuous-wave techniques such as AM- or FMCW.

AMCW Lidar

As with pulsed schemes, AMCW lidars operate by modulating the light's intensity. However, the modulation waveform does not include sharp pulses and carries much less frequency content. Hence, AMCW lidars cannot offer fine range resolution with multi-target detection capability. Nonetheless, the precision of the range measurement can be less than a centimeter, which is sufficient for many applications.

AMCW lidars employ continuous-wave or quasi-continuous-wave laser diodes or LEDs on their transmitter. The intensity can be modulated by varying the bias current of the diode in the electrical domain. The simplicity of these lidars makes them an attractive choice for short-range indoor applications such as gaming and robotics. To reduce the cost of the receiver chip, clever circuit topologies similar to the traditional CMOS imaging pixels have been developed [9, 10].

A simplified circuit schematic and timing diagram of the pixel proposed in [9] are shown in Figs. 4a and 4b. The received light is detected by a single photodiode, but the collected charge is transferred to two separate nodes depending on

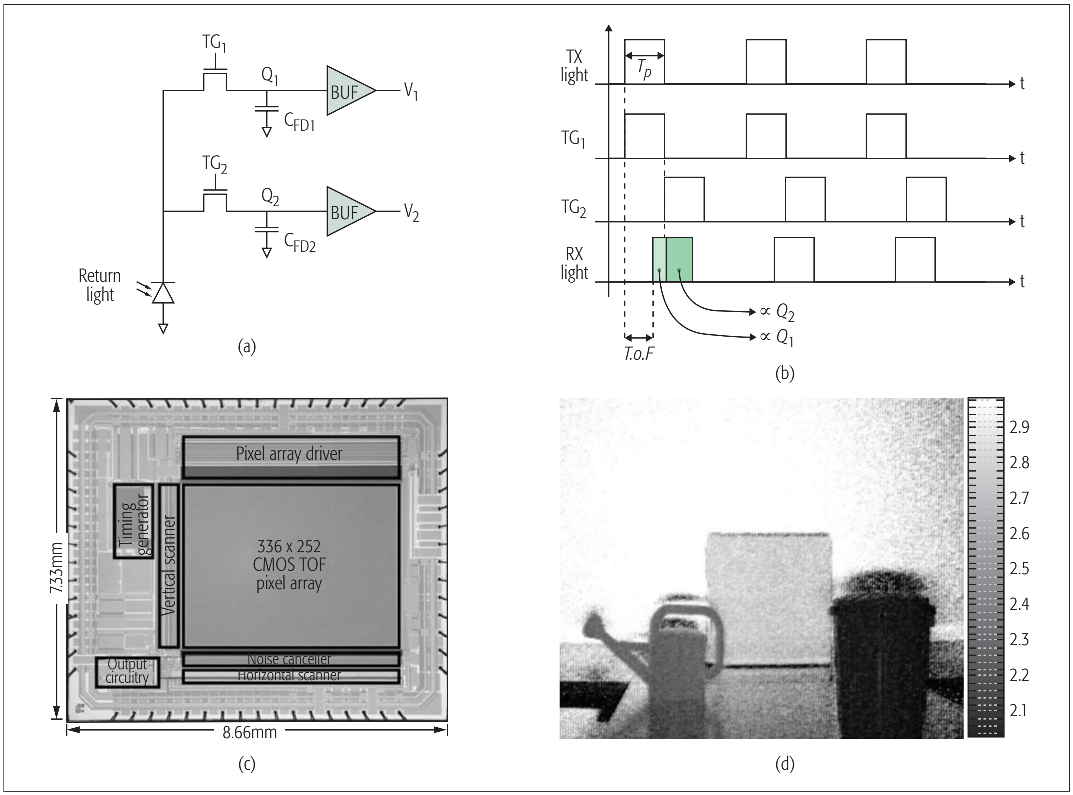


Figure 4. AMCW lidar employed in a CMOS imager for 3D depth measurement in [9]: a) pixel topology; b) timing diagram; c) chip photomicrograph; d) 3D image of scene (scale in meters).

its time of arrival. The charge transfer is controlled by the two transfer gates, TG_1 and TG_2. For short target distances, most of the charge generated by the return light is transferred to node Q_1. For longer distances, the return light experiences more delay, and therefore less charge gets transferred to node Q_1 and more appears at node Q_2. Thus, the ratio of the collected charge at these two nodes is a measure for the time of flight.

The conversion of the time of flight to charge renders this architecture fully compatible with the conventional CMOS RGB pixels that translate the intensity of the ambient light into accumulated charge. Furthermore, the ratiometric nature of the measurement increases its robustness against temperature and process variations and helps suppress the background light and environmental disturbances.

FMCW Lidar

FMCW lidars are fundamentally different from the pulsed and AMCW schemes. Both pulsed and AMCW lidars rely on modulating the intensity of the light. In the receivers of these lidars, the photons are often treated as particles with the range information encoded in their arrival times. In contrast, FMCW lidars rely on the wave properties of the light. In these lidars, the modulation is applied to the frequency of the light field, and an interferometric detection scheme is employed in their receivers [11]. Therefore, the large frequency bandwidth in the optical domain becomes accessible and can be exploited to improve the lidar performance. Unlike pulsed or AMCW lidars, in the FMCW scheme, the interferometric down-conversion of the received signal in the optical domain eliminates the need for wideband electrical circuits. Therefore, mainstream CMOS electronics can be used to achieve exceptional range resolution and precision.

FMCW lidars are fundamentally different from the pulsed and AMCW schemes. Both pulsed and AMCW lidars rely on modulating the intensity of the light. In the receivers of these lidars, the photons are often treated as particles with the range information encoded in their arrival times. In contrast, FMCW lidars rely on the wave properties of the light.

The basic architecture of an FMCW lidar is shown in Fig. 5a. In this case, the frequency of the light emitted from the transmitter is linearly modulated vs. time. The echo light reaches the receiver after the round-trip delay τ_d. For a static target with negligible Doppler effect on the lightwaves, the delay between the collected light and the source causes a constant frequency difference f_d between them, as shown in Fig. 5b. With the linear frequency modulation, $f_d = \gamma \cdot \tau_d$ is directly proportional to τ_d and hence the target range. To measure f_d, a branch of the source light is used as the local oscillator (LO) and is combined with the collected light in a waveguide. The frequency difference between the two light components translates into a periodic phase difference between them and causes an alternating constructive and destructive interference pattern at the frequency f_d. A photodetector is used to convert this pattern into a photocurrent. Measurement of the photocurrent frequency enables range estimation through the following:

$$R = \frac{1}{2} c \cdot \tau_d = \frac{1}{2\gamma} c \cdot f_d \qquad (3)$$

where $\gamma = (\Delta f_{\max})/T$ is the slope of the frequency modulation vs. time with a unit of Hertz per second. This equation demonstrates that the range precision depends on the measurement precision of f_d and also the precision with which the modulation slope γ is controlled or known.

In addition to finer range resolution, the FMCW scheme can also offer much better sensitivity and robustness against environmen-

The mixing gain amplifies the signal before its detection in the photodiode, reducing the electrical noise of the detector referred back to the optical domain. Furthermore, the phase and frequency coherence of the received signal and the LO is necessary to create the interference pattern, rendering the coherent receiver more selective against the ambient light.

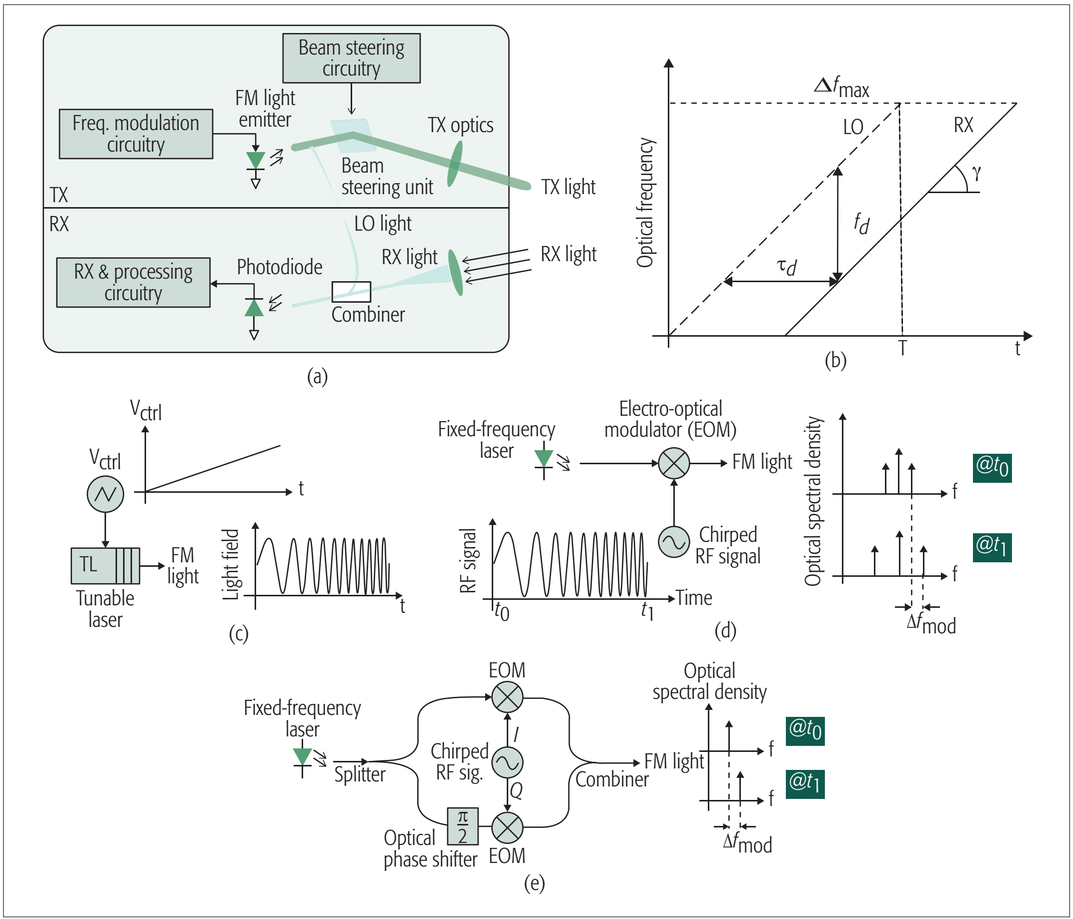


Figure 5. FMCW lidar: a) architecture; b) waveforms. FM light generation using: c) tunable laser; d) electro-optic modulator; e) I/Q modulator.

tal disturbances compared to the pulsed and AMCW lidars because of the FMCW's coherent detection scheme. The interference pattern of the collected beam and the LO in the coherent receiver is similar to the mixing of the two signals in an electrical receiver. The mixing gain amplifies the signal before its detection in the photodiode, reducing the electrical noise of the detector referred back to the optical domain. Furthermore, the phase and frequency coherence of the received signal and the LO is necessary to create the interference pattern, rendering the coherent receiver more selective against the ambient light.

It was previously mentioned that a constant output optical power is desirable for the silicon-photonic-based beam-steering techniques. Therefore, unlike the pulsed architecture, where the large peak power constrains its use in SPPA-based lidars, the fixed light intensity of the FMCW scheme can become increasingly popular as the growing accessibility of SPPAs makes them a mainstream choice for beam-steering lidars.

As illustrated in Figs. 5c and 5d, a tunable laser (TL) or an electro-optic modulator (EOM) can be used to modulate the light's frequency. Tunable lasers are similar to electrical voltage-controlled oscillators (VCOs), but their output is an optical wave rather than an electrical signal. Electro-optic modulators can be viewed as electrical mixers that accept one optical and one electrical signal as their inputs and output an optical signal that is the mix of the two inputs. The frequency of the output optical signal can be tuned by employing a frequency-chirped electrical signal at the modulator input. As with electrical mixers, the electro-optic modulators also create two sidebands in the optical spectrum, as shown in Fig. 5d. In such cases, a coherent receiver capable of detecting both in-phase and quadrature (I/Q) optical fields can be used to extract the target range. An alternative method is to use an I/Q electro-optic modulator to suppress the carrier and create a single-side-band frequency shift in the emitted light [12], as shown in Fig. 5e.

Although both of the aforementioned frequency modulation techniques are theoretically equivalent, there are some practical differences that might make one or the other more suitable for a particular application. The main difference between the two methods is that when using a tunable laser, the frequency tuning happens purely in the optical domain, whereas with an electro-optic modulator, the frequency tuning is generated in the electrical domain and used to modulate the frequency of the light in another step. The modulation bandwidth of a tunable laser can reach beyond 10 THz, which is not achievable by electro-optical modulation. Therefore, architectures based on tunable lasers are more suitable for applications where deep sub-millimeter resolution is necessary.

The possibility of varying a tunable laser's frequency by a large amount and at a fast rate

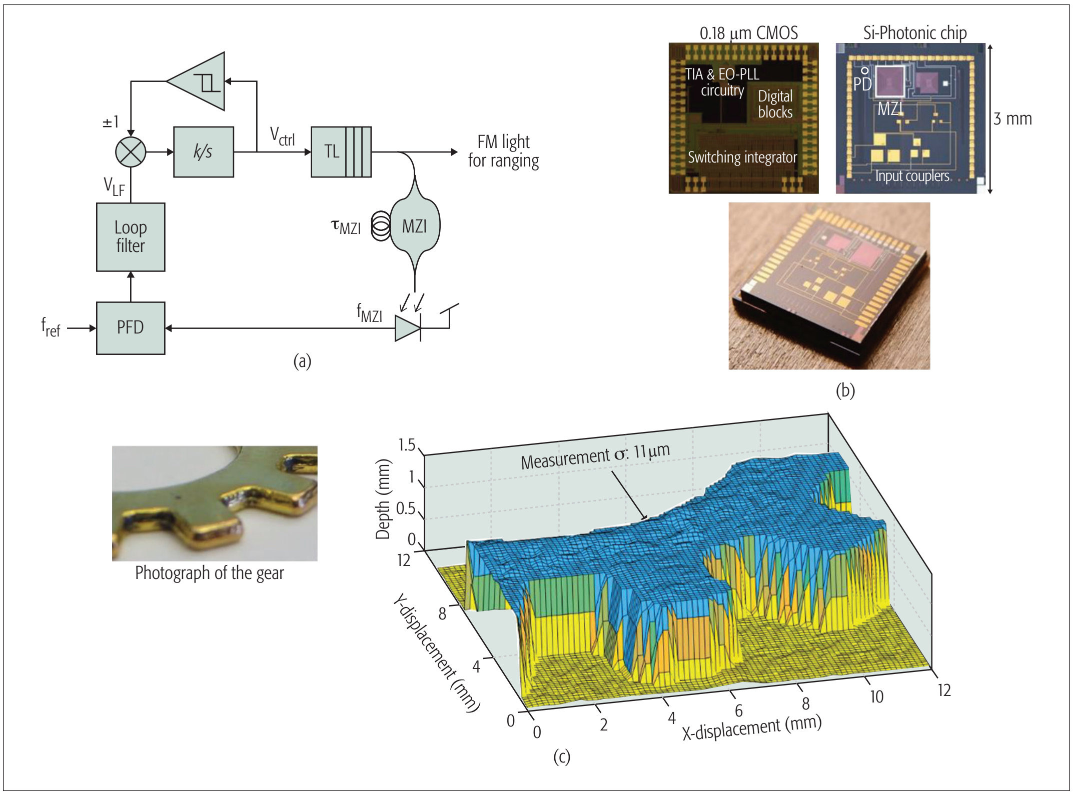


Figure 6. Integrated electro-optical PLL for precision FM light generation a) architecture; b) chip picture and photomicrograph; c) photograph of a gear and its 3D microimage from the FMCW lidar.

The coherence range of widely tunable laser diodes can be as small as a few millimeters, whereas for a fixed-frequency laser employed in an FMCW lidar with electro-optic modulator, the coherence range can reach up to hundreds of meters. This makes the latter a more suitable option for long-range applications.

makes its wavelength more sensitive to noise and environmental disturbances such as temperature variation; hence, widely tunable lasers often suffer from larger phase noise. This phase noise can be tolerated as long as the target range and related delay between the received light from the target and the LO are sufficiently small that the majority of their phase noise is correlated and cancels out in the coherent detection process. However, for long-range lidars, the phase noise of the two light components becomes uncorrelated and the spectrum of the interference signal widens, dropping the power in its fundamental tone. The target range at which the power in the fundamental tone drops to half of its maximum expected value is called the coherence range. This is a measure of the FMCW lidar's maximum operating range. The coherence range of widely tunable laser diodes can be as small as a few millimeters, whereas for a fixed-frequency laser employed in an FMCW lidar with electro-optic modulator, the coherence range can reach up to hundreds of meters. This makes the latter a more suitable option for long-range applications where a few millimeters of resolution is sufficient and wide optical tuning is not needed.

Electronic-Photonic Integrated Circuit for FMCW Lidar

As with a VCO, the frequency of a tunable laser can be controlled in a feedback architecture [13], as illustrated in Fig. 6a. This is achieved by measuring the modulation slope and adjusting it by the laser control signal V_{ctrl} [14]. The modulation slope is measured using a Mach-Zehnder interferometer (MZI), the operation of which is very similar to the FMCW range measurement technique, except the unknown round-trip delay to the target is replaced with a known fixed-length waveguide. Consequently, the interference frequency generated at the output of the MZI is proportional to γ and the waveguide delay: $f_{MZI} = \gamma \cdot \tau_{MZI}$. Because the waveguide delay is fixed, any fluctuations in f_{MZI} can be interpreted as variation in γ. A phase locked loop (PLL) circuit can be used to measure these fluctuations against a reference frequency f_{ref}, and the fluctuations can be suppressed by adjusting the laser control signal V_{ctrl} to ensure that $\gamma = f_{ref}/\tau_{MZI}$. Because the linear modulation cannot continue indefinitely, a hysteresis comparator observes the level of and reverses its slope (to generate up/down ramps) whenever it crosses some predefined boundaries.

The control loop for the laser modulation is implemented on a heterogeneously integrated electronic-photonic chip stack as described in [13]. The MZI and the photodetector are fabricated on a silicon-photonic chip, and the electronic circuits are designed in a 0.18 m CMOS process. The two dies are integrated using through-silicon-vias (TSVs) to make a single chip-stack as shown in the photograph.

This electronic-photonic integrated circuit modulates the frequency of a discrete tunable laser with high precision and repeatability. The output light of the laser is used to create a 3D image of a gear placed at a 40-cm distance from the source, at a rate of 10 kP/s with 11-m range precision and 250-m lateral resolution. The 3D image reconstructed from this measurement is shown in Fig. 6c.

While the objective of this particular work was to achieve a fine range precision, the design trade-offs explained herein can also be used to guide the development of FMCW lidars that are more suitable for long-range applications with lower range resolution [15].

Conclusion

Accurate detection of the surrounding environment is of the utmost importance to the successful operation of autonomous machines such as self-driving cars, drones, and indoor robots. Among different sensory systems, 3D cameras have proven to be an essential aid for such machines, providing precise dimensions of and distances to objects in their vicinity. Among different 3D imaging techniques, the fine volumetric resolution and long operational distance of lidar-based solutions have significantly surpassed those of other techniques. Many different lidar architectures have been investigated over the last several decades. Among them, FMCW lidars provide the finest resolution for short-range applications. Because of their coherent detection scheme, they can also detect the lowest returning light levels from distant targets at the fundamental shot noise limit. In addition, the constant optical power level at their output is compatible with the emerging silicon-photonic-based optical phased arrays for beam steering, which cannot easily accommodate the large peak power of a pulsed lidar. This is particularly important, because the high cost of the current beam steering solutions is one of the major challenges in developing inexpensive long-range lidars, and silicon-photonic phased-array is one of the most promising technologies that can solve this issue. These characteristics have made the FMCW lidars an increasingly attractive choice for applications from those in advanced medical and scientific fields to self-driving cars and drones. As photonic devices become more accessible through monolithic CMOS processes or heterogeneously integrated silicon-photonic and CMOS platforms, more flavors in FMCW transmit and receive architectures will lead to fully integrated next-generation lidars that can be designed and optimized for a wide range of applications.

As photonic devices become more accessible through monolithic CMOS processes or heterogeneously integrated silicon-photonic and CMOS platforms, more flavors in FMCW transmit and receive architectures will lead to fully integrated next-generation lidars that can be designed and optimized for a wide range of applications.

Acknowledgment

The authors would like to thank Dr. Niels Quack and Dr. Tae-Joon Seok, former members of the lidar research team at the University of California at Berkeley, and also Dr. Mahdi Kashmiri and Dr. Ken Wojciechowski from Robert Bosch LLC Research and Technology Center in Palo Alto for their critical feedback on this article. The authors would also like to thank DARPA's Diverse Accessible Heterogeneous Integration Program as the main funding agency for this research; the TSMC University Shuttle Program for CMOS chip fabrication; the Marvell Nanofabrication Laboratory at the University of California at Berkeley for the fabrication of the silicon-photonic chips; and the Berkeley Sensor and Actuator Center and the UC Berkeley Swarm Lab for their continuous support.

References

[1] R. Halterman and M. Bruch, "Velodyne HDL-64E LIDAR for Unmanned Surface Vehicle Obstacle Detection," *Proc. SPIE Defense, Security, and Sensing, Int'l. Society for Optics and Photonics*, Orlando, FL, 2010.

[2] E. S. Cameron, R. P. Szumski and J. K. West, "Lidar Scanning System," U.S. Patent 5006721, 9 Apr. 1991.

[3] D. P. Resler *et al.*, "High-Efficiency Liquid-Crystal Optical Phased-Array Beam Steering," *Optics Letters*, vol. 21, no. 9, 1996, pp. 689–91.

[4] B.-W. Yoo *et al.*, "A 32 32 Optical Phased Array Using Polysilicon Sub-Wavelength High-Contrast-Grating Mirrors," *Optics Express*, vol. 22, no. 16, 2014, pp. 19,029–39.

[5] J. Sun *et al.*, "Large-Scale Nanophotonic Phased Array," *Nature*, vol. 493, no. 7431, 2013, pp. 195–99.

[6] K. Van Acoleyen *et al.*, "Off-Chip Beam Steering with a One-Dimensional Optical Phased Array On Silicon-On-Insulator," *Optics Letters*, vol. 34, no. 9, 2009, pp. 1477–79.

[7] P. M. Woodward, *Probability and Information Theory, with Applications to Radar*, Pergamon Press, 1953.

[8] C. Niclass *et al.*, "Design and Characterization of a CMOS 3-D Image," *J. Solid-State Circuits*, vol. 40, no. 9, 2005, pp. 1847–54.

[9] S. Kawahito *et al.*, "A CMOS Time-of-Flight Range Image Sensor with Gates-on-Field-Oxide Structure," *IEEE Sensors J.*, vol. 7, no. 12, 2007, pp. 1578–86.

[10] W. Kim *et al.*, "A 1.5Mpixel RGBZ CMOS Image Sensor for Simultaneous Color and Range Image Capture," *Proc. Int'l. Solid-Sate Circuits Conf.*, San Francisco, CA, 2012.

[11] D. Uttam and B. Culshaw, "Precision Time Domain Reflectometry in Optical Fiber Systems Using a Frequency Modulated Continuous Wave Ranging Technique," *J. Lightwave Technology*, vol. 3, no. 5, 1985, pp. 971–77.

[12] P. A. Sandborn *et al.*, "Dual-Sideband Linear FMCW Lidar with Homodyne Detection for Application in 3D Imaging," *Proc. Conf. Lasers and Electro-Optics*, San Jose, CA, 2016.

[13] B. Behroozpour *et al.*, "Electronic-Photonic Integrated Circuit for 3D Microimaging," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, 2017, pp. 161–72.

[14] N. Satyan *et al.*, "Precise Control of Broadband Frequency Chirps Using Optoelectronic Feedback," *Optics Express*, vol. 17, no. 18, 2009, pp. 15991–99.

[15] G. N. Pearson *et al.*, "Chirp-Pulse-Compression Three-Dimensional Lidar Imager with Fiber Optics," *Applied Optics*, vol. 44, no. 2, 2005, pp. 257–65.

Biographies

Behnam Behroozpour [M'16] received his B.Sc. degree from Sharif University of Technology, Tehran, Iran, in 2010, his M.Sc. degree from the University of Twente, Enschede, the Netherlands, in 2012, and his Ph.D. degree from the University of California at Berkeley in 2016. He is currently a research engineer with Bosch LLC, Palo Alto, California. His research interests include electronic and photonic integrated circuits, MEMS, LIDAR, and 3D imaging technologies.

Phillip A.M. Sandborn received his B.Sc. degree in electrical engineering and mathematics from the University of Maryland, College Park, in 2012. He is currently pursuing a Ph.D. degree in electrical engineering with the University of California at Berkeley. His current research interests include 3D imaging using LIDAR, optical phase-locked loops, optical packaging, and noise-limited performance of optical systems.

Ming C. Wu [F'02] received his Ph.D. degree from the University of California at Berkeley in 1988. He is currently a Nortel Distinguished Professor of Electrical Engineering and Computer Sciences with the University of California at Berkeley. He was a Packard Foundation Fellow from 1992 to 1997. He received the 2007 Paul F. Forman Engineering Excellence Award from the Optical Society of America and the 2016 William Streifer Award from the IEEE Photonics Society.

Bernhard E. Boser [F'03] received his Ph.D. degree from Stanford University, California, in 1988. In 1992, he joined the Faculty of the EECS Department, University of California at Berkeley. He is also a co-founder of SiTime, Santa Clara, California, and Chirp Microsystems, Berkeley, California. He served as the President of the Solid-State Circuits Society and on the Program Committees of ISSCC, VLSI Symposium, and Transducers. He also served as the Editor-in-Chief of the *IEEE Journal of Solid-State Circuits.*

Integrated Circuit Technology for Next Generation Point-of-Care Spectroscopy Applications

Jonas Handwerker, Benedikt Schlecker, Maurits Ortmanns, and Jens Anders

ABSTRACT

Point-of-care personalized medicine and home diagnostics are emerging topics that can help to master the challenges of aging societies. Magnetic resonance spectroscopy is one of the most promising sensing principles because it enables the detection of proteins, metabolites, and reactive oxygen species, which play crucial roles in a large number of diseases, with high specificity. To provide a self-contained introduction to the topic, this article starts with a description of the basic working principle of magnetic resonance spectroscopy, highlighting the similarities and differences compared to conventional impedance spectroscopy methods. Focusing on the two specific techniques of NMR and ESR spectroscopy, we explain how miniaturized systems co-integrating detectors and the signal processing electronics on a single chip are a key enabler for portable, low-cost spectrometry systems. These systems bear many similarities to conventional communication transceivers and can therefore largely benefit from recent advances in communication circuits as well as entirely new detection principles such as VCO-based detection, which are enabled by the use of modern nanometer-scaled integrated circuit technologies. An overview of the current state of the art of such miniaturized magnetic resonance spectrometers is presented, which both highlights the excellent new possibilities associated with these systems and at the same time outlines the current challenges and future research directions in this emerging field of research.

INTRODUCTION

Spectroscopic techniques play an important role in a large number of disciplines ranging from analytical chemistry over materials science to medical and biomedical applications. Especially in the medical sector, with recent advances in sensor technologies for point-of-care (PoC) applications and ubiquitous data accessibility thanks to the Internet of Things (IoT), the fields of personalized medicine and home diagnostics have gained tremendous interest. Here, among other sensing principles, impedance spectroscopy has emerged as one of the prime candidates for future PoC diagnostic applications because it can greatly benefit from the excellent performance, the miniaturization potential, and the low costs for high volume production associated with modern nanometer-scaled integrated circuit (IC) technologies. Since the frequency range used for impedance spectroscopy coincides with the radio frequency band (3 kHz–300 GHz) and the generic architecture of an impedance spectrometer bears many similarities to a standard transceiver for communication applications, these spectroscopy applications present an interesting new playground for researchers from electronics and communications engineering.

The intention of this tutorial article is to both give the interested novice a self-contained introduction into the topic and at the same time provide the expert with an overview of recent developments and future trends in the field. To this end, we use example IC-based realizations, recently published in the literature, which highlight how transceiver architectures for modern communications applications with their advanced on-chip functionality can positively impact the field of PoC spectroscopy.

DIELECTRIC SPECTROSCOPY

In the most general sense, spectroscopic methods exploit the interaction between electromagnetic (EM) fields and matter to characterize the sample under investigation. Today, with the availability of advanced detection hardware, the entire EM spectrum from frequencies in the subhertz region to very high frequencies exceeding petahertz can be used (Fig. 1a). Naturally, the type of interaction and therefore also the type of spectroscopic information that can be studied greatly varies with the utilized frequency band. This is illustrated in Fig. 1b, which shows the different types of direct resonant interaction mechanisms that occur between an EM field and matter. Here, resonant refers to the fact that the interaction in principle occurs at a single frequency, resulting in a spectral peak that can be related to said resonant absorption for spectroscopic purposes. The advantage of a resonant interaction is that it intrinsically provides much greater specificity than a non-resonant one.

According to Fig. 1a, for frequencies below approximately 300 GHz, which we consider the usable frequency range for all-electronic spectrometers, dielectric spectroscopy can be used to probe a sample under investigation. Following Fig. 1b, in this frequency range, the only direct resonant interaction between an EM field and a

Point-of-care personalized medicine and home diagnostics are emerging topics that can help to master the challenges of aging societies. Magnetic resonance spectroscopy is one of the most promising sensing principles because it enables the detection of proteins, metabolites, and reactive oxygen species, which play crucial roles in a large number of diseases, with high specificity.

The authors are with the University of Ulm.

Digital Object Identifier: 10.1109/MCOM.2017.1700092

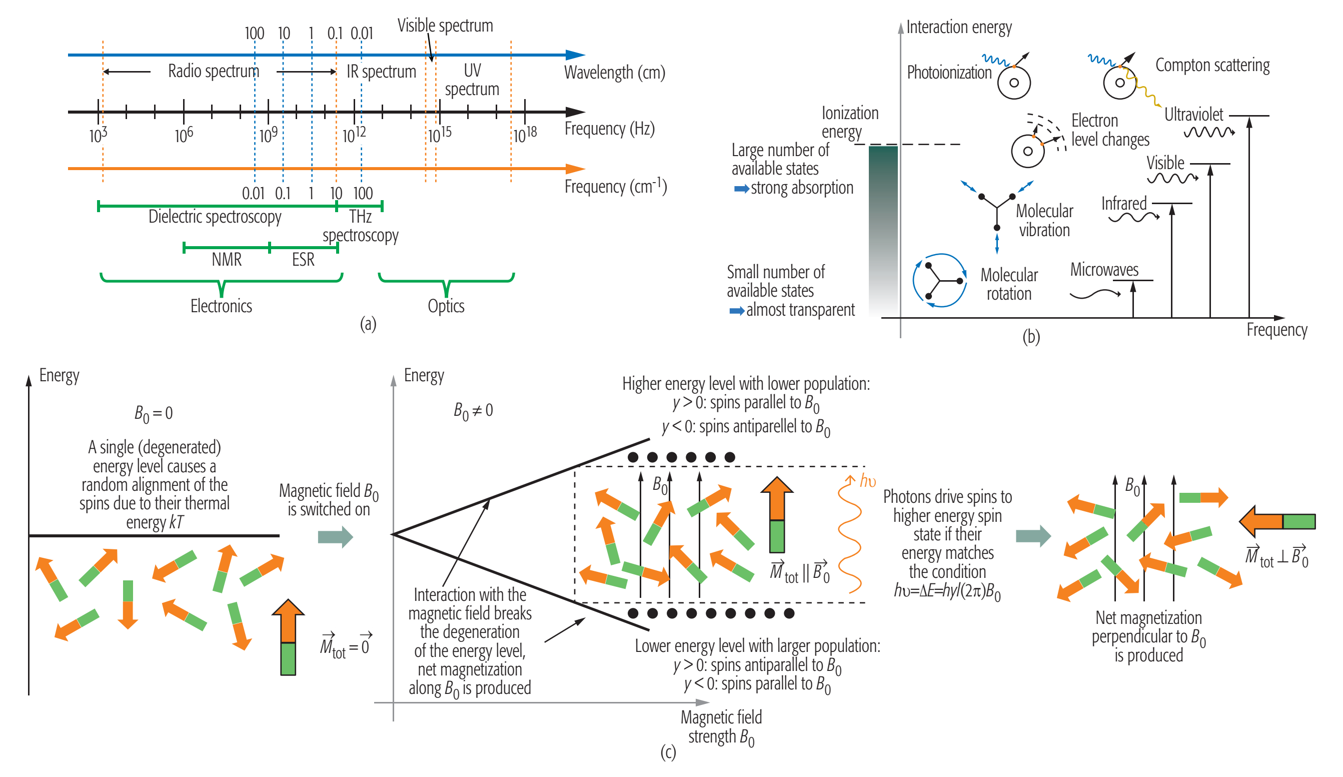

100 10 1 0.1 0.01
Visible spectrum
Radio spectrum
IR spectrum
UV spectrum
Wavelength (cm)
10^3 10^6 10^9 10^{12} 10^{15} 10^{18}
Frequency (Hz)
0.01 0.1 1 10 100
Frequency (cm^{-1})
Dielectric spectroscopy
THz spectroscopy
NMR
ESR
Electronics
Optics
(a)
Interaction energy
Photoionization
Compton scattering
Ionization energy
Electron level changes
Ultraviolet
Large number of available states
➡strong absorption
Visible
Molecular vibration
Infrared
Small number of available states
➡almost transparent
Molecular rotation
Microwaves
Frequency
(b)
Energy
$B_0 = 0$
A single (degenerated) energy level causes a random alignment of the spins due to their thermal energy kT
$\vec{M}_{tot} = \vec{0}$
Magnetic field B_0 is switched on
Energy
$B_0 \neq 0$
Higher energy level with lower population:
$\gamma > 0$: spins parallel to B_0
$\gamma < 0$: spins antiparallel to B_0
B_0
$h\upsilon$
$\vec{M}_{tot} \parallel \vec{B_0}$
Interaction with the magnetic field breaks the degeneration of the energy level, net magnetization along B_0 is produced
Lower energy level with larger population:
$\gamma > 0$: spins antiparallel to B_0
$\gamma < 0$: spins parallel to B_0
Magnetic field strength B_0
Photons drive spins to higher energy spin state if their energy matches the condition
$h\upsilon = \Delta E = h\gamma/(2\pi)B_0$
B_0
$\vec{M}_{tot} \perp \vec{B_0}$
Net magnetization perpendicular to B_0 is produced
(c)

Figure 1. a) EM spectrum highlighting the frequency ranges of the spectroscopic methods discussed in this article; b) illustration of the different direct mechanisms by which EM fields can interact with matter; c) illustration of magnetic resonance spectroscopy, where an externally applied static magnetic field introduces sharp resonant interactions in the radio frequency band.

sample occurs due to molecular rotations, and the corresponding spectroscopy is called rotational spectroscopy. Rotational spectroscopy is widely used to investigate fundamental aspects of molecular physics of gas phase molecules. The generic setups for continuous-wave (cw) and Fourier transform (FT) dielectric spectroscopy experiments are shown in Figs. 2a and b. The two methods differ in the way the sample is excited to extract the spectroscopic information in the form of the complex permittivity $\varepsilon = \varepsilon' - j\varepsilon''$: In cw experiments, the sample is continuously irradiated, and the reflected power of the sample, which is placed inside a resonator to enhance sensitivity, is measured as a function of frequency to extract ε. In FT experiments, ε is extracted by exciting the sample with a short (i.e., broadband) pulse and measuring the transmission through the sample. Here, in both cases, the complex permittivity ε can conveniently be obtained from the outputs of a conventional quadrature receiver.

All remaining resonant interactions of Fig. 1b occur for frequencies that are not readily accessible using all-electronic setups and require optical or even ultraviolet and X-ray detectors. Therefore, although these techniques can also largely benefit from IC realizations of the required electronics in the forms of systems-in-a-package or even systems-on-a-chip (e.g., [1, 2]), they are not further discussed in this article, which focusses on all-electronic PoC spectroscopy systems.

Magnetic Resonance Spectroscopy

An alternative approach to introduce the desired resonant absorption in the radio frequency range, which can be covered by purely electronic detection systems, exploits the magnetic resonance (MR) effect of a spin ensemble associated with the sample under investigation and the applied EM field. In essence, this approach utilizes the fact that both electron and nuclear spins are associated with a magnetic moment, which can interact with externally applied magnetic fields. Without going into the quantum mechanical details, the gist of the MR method is illustrated in Fig. 1c. Here, the key observation is that the desired resonant interaction is introduced by an externally applied magnetic field B_0, which splits the degenerated (at $B_0 = 0$) energy level associated with the spin angular momentum into — for spin-half particles such as protons and electrons — two distinct energy levels (Zeeman effect) [3]. The resulting energy gap between the two energy levels depends on the gyromagnetic ratio γ of the particle. For protons, as they are used in ^{1}H-nuclear magnetic resonance (NMR) spectroscopy, γ takes a value of $\gamma_{1H} \approx 2\pi \cdot 42$ MHz/T. For electrons, as they are observed in electron spin resonance (ESR) spectroscopy, γ is given by $\gamma_{e^-} \approx -2\pi \cdot 28$ GHz/T. As a consequence, assuming standard magnetic field strengths between ≈ 0.1 T and ≈ 23 T, NMR resonances occur in the frequency band between a few megahertz and approximately 1 GHz, while ESR resonances cover the spectral range between a few gigahertz and approximately 500 GHz. Since macroscopically, the magnetic resonance effect manifests itself as a change of the effective sample susceptibility $\mu = \mu' - j\mu''$, the experimental setups are very similar to those for generic impedance spectroscopy shown in Fig. 2. In fact, the experimental setup for so-called cw experiments is (up to the need for

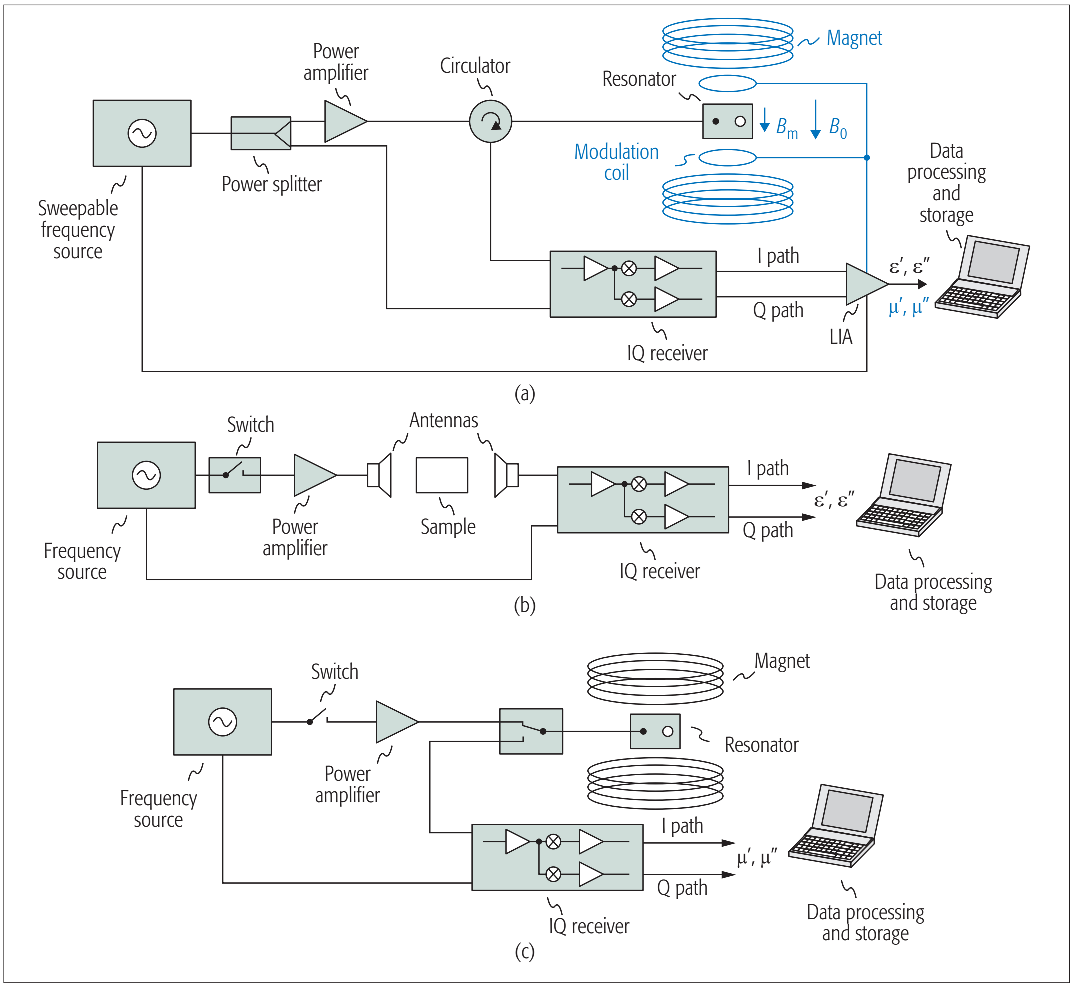


Figure 2. a) Generic setup for continuous-wave spectroscopy experiments; components in blue are additionally needed for cw magnetic resonance experiments; b) generic setup for Fourier transform (FT) impedance spectroscopy experiments; c) generic setup for FT (also called "pulsed") magnetic resonance spectroscopy experiments.

The spectroscopic information obtained from both NMR and ESR is very rich because the magnetic moments associated with both nuclear and electron spins effectively act as magnetic nanoprobes inside a molecule, which are very sensitive to their magnetic and electronic environment.

the static magnetic field generation) identical to the dielectric spectroscopy setup. The setup of FT or pulsed magnetic resonance experiments is simplified by the fact that the spin ensemble not only produces a measurable signal during the pulse but also after a pulsed excitation. This fact allows for a reflection-type pulsed measurement according to Fig. 2c, resulting in a simpler experimental setup compared to pulsed dielectric spectroscopy. Moreover, in pulsed MR experiments, dedicated sequences of different excitation pulses can be used to greatly enrich the available spectroscopic information, and MR pulse sequence research is a very active field.

The spectroscopic information obtained from both NMR and ESR is very rich because the magnetic moments associated with both nuclear and electron spins effectively act as magnetic nanoprobes inside a molecule, which are very sensitive to their magnetic and electronic environment. Therefore, the resulting small resonance frequency shifts provide detailed information about the structure of the molecule containing the spin, turning NMR and ESR into two of the most powerful spectroscopic techniques available today. This being said, their major drawbacks, which prevent wide use of low-cost, in-field, PoC spectroscopy applications, are the relatively poor sensitivity as well as the large instrument size and cost. Fortunately, thanks to their excellent miniaturization capabilities and low costs at high-volume production, application-specific integrated circuit (ASIC)-based approaches are ideally suited to solve the latter problem especially, and the first existing academic prototypes of ASIC-based NMR and ESR spectrometers show very promising results. Therefore, NMR and ESR spectrometry can be considered as two of the prime candidates for future generations of portable PoC analysis systems.

IC-Based Magnetic Resonance Point-of-Care Spectrometry

In the following two sections, we discuss NMR and ESR separately concerning their specific requirements and advantages as well as disadvantages for the realization of ASIC-based PoC spectrometers.

NMR Spectroscopy

NMR exploits the interaction between the intrinsic magnetic moment of nuclei with odd numbers of nucleons and an external magnetic field B_0. The resonance frequency, which in NMR and ESR terminology is called the Larmor frequency, ω_L =

The frequency range and circuit architecture of a pulsed NMR experiment are very similar to a classical transceiver for communication purposes, rendering NMR an interesting new application field that can substantially benefit from architectural and circuit-level innovations in electronics for communication systems.

$-\gamma \cdot B_0$, is directly proportional to the strength of the externally applied magnetic field, and the proportionality constant γ is the gyromagnetic ratio mentioned above. The proportionality between ω_L and B_0 can, for example, be used for NMR-based imaging by making ω_L a function of position using gradient fields.

Due to the high hydrogen concentration in most molecules, ^{1}H is by far the most frequently used NMR nucleus today. However, thanks to improvements in detector sensitivity, NMR spectroscopy on different nuclei such as ^{13}C, ^{23}Na, ^{31}P, and ^{39}K is receiving continuously growing attention. These nuclei are particularly interesting for medical applications where they provide additional information for the analysis of proteins, metabolites, and neurotransmitters and their role in cellular processes. The gyromagnetic ratios differ substantially for different nuclei, for example, $\gamma_{1H} \approx 2\pi \cdot 42.6$ MHz/T and $\gamma_{39K} \approx 2\pi \cdot 1.99$ MHz/T, turning the simultaneous excitation and readout of NMR spectra associated with different nuclei into a difficult task. The spectroscopic information in NMR is encoded in small frequency shifts induced by the magnetic and electronic interaction of the spin with its environment in the molecule. Here, these shifts are typically in the low ppm range, resulting in maximum absolute NMR-induced frequency shifts below tens of kilohertz. Since, additionally, the relaxation times encountered in NMR range from tens of microseconds for solid up to several seconds for liquid samples, it is possible to excite the entire NMR spectrum corresponding to a single nucleus with a single pulse with a duration in the microsecond range without significant relaxation effects during the pulse. Today, such pulses can easily be generated using standard hardware, and virtually all NMR experiments are performed in the pulsed mode according to Fig. 2c.

According to Figs. 1a and 2c, the frequency range and circuit architecture of a pulsed NMR experiment are very similar to a classical transceiver for communication purposes, rendering NMR an interesting new application field that can substantially benefit from architectural and circuit-level innovations in electronics for communication systems. The main difference from a standard architecture for communication lies in the fact that the NMR transceiver electronics are connected to a coil instead of an antenna. This is because, in contrast to most communication systems, where the antenna produces an outgoing/receives an incoming EM wave, in NMR, the coil operates in its near-field. That is, it excites the spins by the RF-magnetic field, the so-called B_1 field, produced by the current running through its windings, and it senses the spins as an induced electromotive force (emf). Depending on the desired hardware complexity, NMR setups can use either separate coils for sample excitation (TX) and signal detection (RX) or a single coil in transmit-receive (TX/RX) mode. Here, a separate realization offers the advantage of an individual optimization of the coils and their tuning circuits for each mode. In contrast, using a single TX/RX coil in combination with a TX/RX switch simplifies the overall experimental setup.

Since conventional NMR setups are constructed using discrete electronics, providing limited possibilities for sophisticated digitally assisted calibration schemes, there is, in principle, large room for improvements in the NMR experimental setup by using monolithic complementary metal oxide semiconductor (CMOS) transceiver realizations with their greatly improved design freedom. Consequently, over the last years, a number of different designs using both integrated [4-6] and external [6-8] NMR coils have been presented in the literature, which exploit the potential of IC-based transceiver realizations. Here, one of the most important advantages of IC-based transceiver realizations is the possibility of placing the electronics in close proximity to the NMR coil, removing the need for a 50 Ω-impedance matching between coil and IC as is required for a remote transceiver placement with a 50 Ω-transmission line between coil and transceiver chip. An important consequence is that the resulting freedom to choose the impedance can be used for optimized transmission efficiency and/or detection sensitivity.

Here, the TX efficiency is optimized by maximizing the coil current to maximimize the B_1 field produced by the coil. Assuming a simple H-bridge power amplifier (PA), [4, 6], this is achieved by minimizing the effective impedance, for example, by embedding the coil into a series resonant LC circuit. Interestingly, due to their small size and correspondingly low inductance, integrated and external TX microcoils can be directly driven by simple CMOS H-bridge PAs without the need for any impedance matching network, still achieving $\pi/2$ excitation pulse lengths around 10 μs [4]. For non-shaped pulses, the PA linearity is of less importance because the spins can only be resonantly exited at the Larmor frequency, and harmonics only have a very small effect on the NMR experiment. However, advanced NMR spectroscopy and relaxometry experiments require the ability to modify both the phase and the amplitude of the transmit pulse.

In the RX path, the induced NMR emf is proportional to the B_1 field normalized to the coil current I_{coil}, $B_u = B_1/I_{coil}$, and is therefore independent of I_{coil}. The overall receiver noise figure (NF) can be improved by embedding the RX coil into a parallel LC resonant circuit, which provides a noise-free preamplification of both the NMR voltage and the coil noise by the quality factor of the LC tank. This in turn relaxes the noise requirements of the subsequent low noise amplifier (LNA). In this case, the LNA should be designed as a voltage amplifier with a large input impedance, which can most easily be designed in CMOS technology. Due to the intrinsically differential nature of the induced emf, differential LNA realizations are preferred [4]. In contrast to classical LNAs for communication purposes, which always have to trade off noise performance vs. linearity, LNAs for NMR applications can frequently be designed with a pure focus on the noise performance. The linearity is less important unless a large NMR peak, for example, produced by a solvent of high concentration, typically water, is present in the spectrum along with very small peaks from lower-concentration molecules.

Since the design goals for TX and RX are usually contradictory, and the quality factors of multi-turn integrated coils are typically relatively low (10

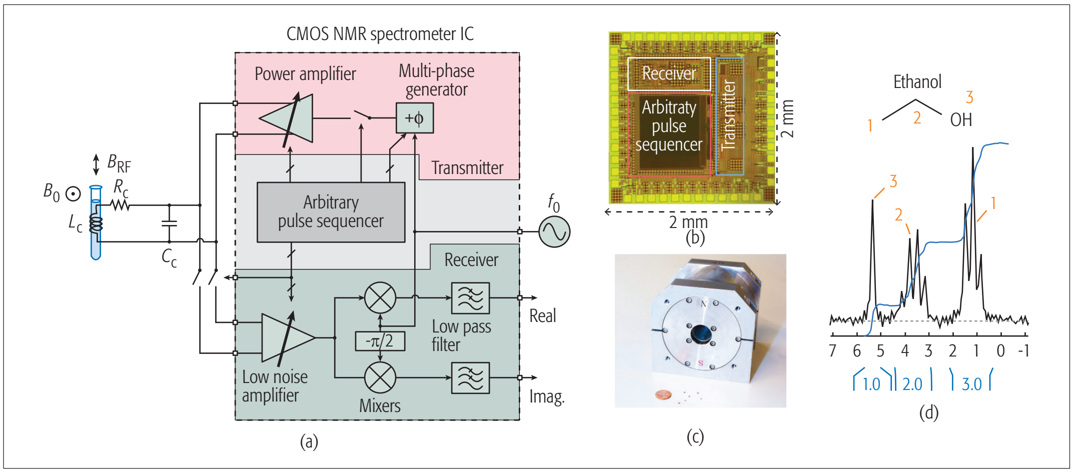


Figure 3. NMR spectrometer according to [8]: a) chip architecture containing a quadrature receiver and a power amplifier with phase- and amplitude modulation capabilities; b) chip micrograph of the transceiver ASIC with a chip area of 2 × 2 mm^2; c) photo of the 0.51 T NdFeB permanent magnet (W × D × H: 12.6 × 11.7 × 11.9 cm^3, weight: 7.3 kg, Neomax Co.); d) 1D NMR spectrum of Ethanol recorded in a single scan.

Despite the many advantages of a monolithic integration of the ESR spectrometer electronics, its major limitation lies in the limited available output power at the elevated ESR frequencies provided by standard IC technologies such as CMOS and BiCMOS.

or even below), very recently, a couple of designs have been presented that omit impedance matching altogether and directly connect the (untuned) coil to the transceiver [5, 6]. One of the major advantages of the untuned excitation and readout is that it allows for very broadband operation, which in turn provides a very elegant and efficient possibility for the detection of a large number of different nuclei with the same front-end (X-nuclei NMR) with the associated richer spectroscopic information (e.g., [6]).

Most integrated NMR transceivers use quadrature mixers with their improved noise performance and increased IF bandwidth to demodulate the detected signal to the baseband. For low-field PoC systems, which are typically operated below 0.5 T, corresponding to an operating frequency < 21 MHz for ^1H NMR, direct digitization of the LNA output is an alternative approach that can help to reduce system complexity and offers more flexibility by digital signal processing (e.g., [9]).

As a first example of an IC-based PoC NMR spectroscopy system, we discuss the compact NMR spectroscopy system presented in [8]. The design uses a fully integrated transceiver realized in a 0.18 μm CMOS technology in combination with an off-chip solenoidal TX/RX coil inside a 0.51 T permanent magnet corresponding to an operating frequency of 21.8 MHz for ^1H (Figs. 3a–3c). The system is capable of generating a broad variety of excitation pulses, which in turn enable a large variety of NMR experiments, including conventional 1D NMR spectroscopy as well as 2D NMR spectroscopy and NMR relaxometry. The homogeneity of the utilized permanent magnet is not sufficient for the analysis of large molecules such as proteins, but enables PoC spectroscopy for small and medium-size bio-relevant molecules. An example spectrum of ethanol recorded with the IC-based system is shown in Fig. 3d.

The recorded spectrum is an instructive example of the capabilities of NMR spectroscopy. In Fig. 3d, the red numbers 1 to 3 indicate the three different types of protons in the CH_3, CH_2, and OH groups of the ethanol molecule, respectively. From the location of the main peaks and their J-coupling induced splitting in the NMR spectrum, the molecular structure of the ethanol molecule can be deduced. Moreover, NMR is even a quantitative spectroscopic technique because the integral over each peak is directly proportional to the number of spins contributing to said peak, as indicated in blue with a ratio of 3:2:1.

In [5], a PoC NMR relaxometry system with an even higher level of integration was presented (Fig. 4). The design is based on an NMR transceiver realized in a 0.18 μm CMOS technology and optimized for the operation inside a 0.46 T magnet. The system operates from a battery supply without the need for any external instruments except a computer for data analysis, display, and storage. The design features an untuned 23-turn planar on-chip TX/RX coil with an area of 2.0 × 2.0 mm^2. Since the homogeneity of the utilized magnet is worse compared to [8], only NMR relaxometry experiments using functionalized magnetic nanoparticles (e.g., for the detection of labeled DNA or proteins) were presented.

ESR Spectroscopy

In ESR spectroscopy, the resonance produced by the interaction between the spin of an electron and an externally applied magnetic field B_0 is exploited to obtain the spectroscopic information. Therefore, only substances that possess a nonzero net electron spin (paramagnetic substances, free radicals) can be directly analyzed by ESR spectroscopy. Although this might sound like a major limitation, there are a large number of important material classes that are ESR active. To name just one example, reactive oxygen species (ROS) play a major role in the development of many diseases including cancer, Alzheimer's disease, atherosclerosis, autism, and Parkinson's disease, and ESR presents the gold standard for their detection. Moreover, in addition to substances that can be directly measured using ESR, the growing fields of ESR spin labeling, spin trapping, and spin probing also make molecules accessible to ESR-based analysis, which are intrinsically ESR silent or have lifetimes that make their direct ESR-based measurement difficult.

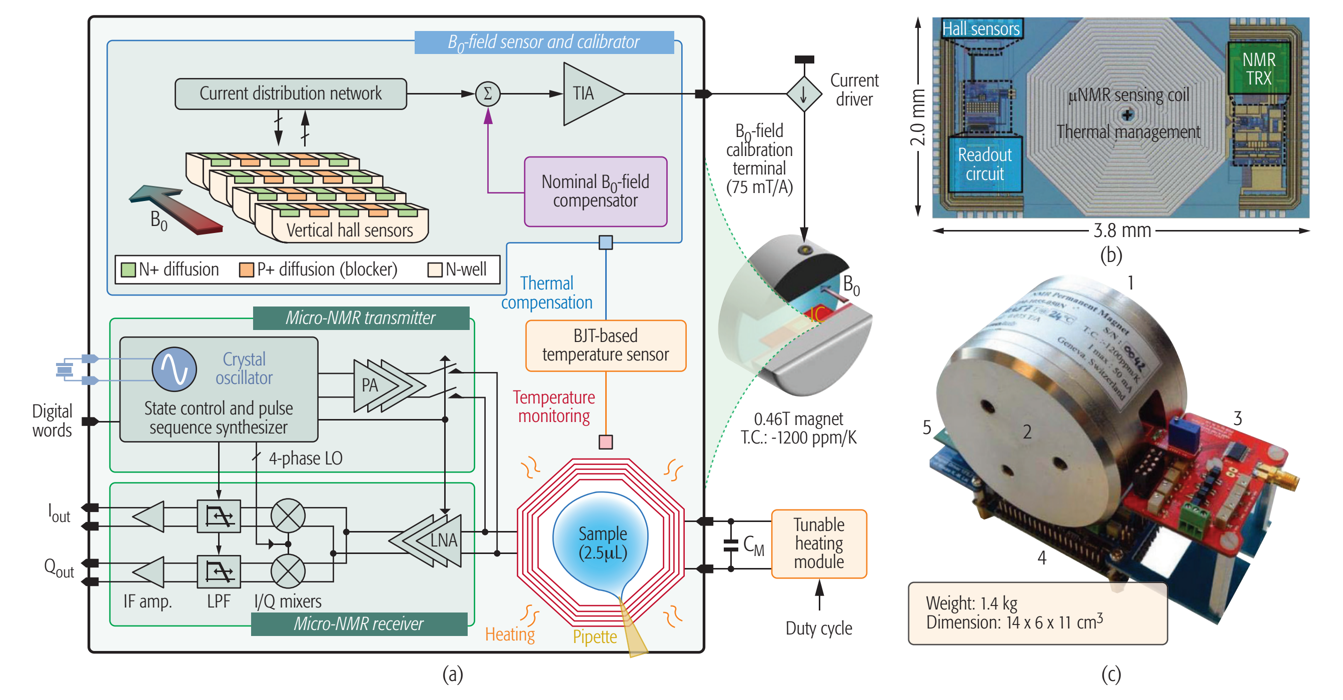


Figure 4. NMR relaxometry system presented in [5]: a) system and chip architecture containing a quadrature receiver, a power amplifier with phase modulation capabilities, and a Hall-sensor for field-drift compensation; b) chip micrograph of the transceiver ASIC with integrated TX/RX coil and a chip area of 2.0 × 3.8 mm^2; c) photo of the battery-operated system containing the ASIC, an FPGA, and a current driver for field compensation (magnet 0.46 T NdFeB, weight 1.4 kg, METROLAB Instruments SA).

One major challenge of pulsed ESR experiments (Fig.2c) compared to pulsed NMR experiments is related to the significantly shorter relaxation times of electron spins (nanoseconds to microseconds) compared to nuclear spins. Without going into detail, these short relaxation times impose two major constraints on ESR:

- In the transmit path, the excitation pulse has to be shorter than the sample's relaxation time to avoid relaxation during the pulse. The resulting short pulse lengths (nanosecond range) in turn mandate large pulse powers at very elevated frequencies to achieve the $\pi/2$ and even π flip angles required by many pulse sequences. This is because the flip angle θ is related to the pulse length τ and the excitation power PTX according to

$$\theta \propto \gamma \cdot \tau \cdot \sqrt{P_{\mathrm{TX}}}.$$

Moreover, the excitation pulse needs to be short enough to excite a sufficiently large portion of the spectrum — note that the excited spectral range is proportional to $1/\tau$. Since the interactions encountered in ESR spectroscopy lead to spectra that cover a much larger frequency range than those encountered in NMR, frequently, cw experiments still provide the desired spectroscopic information in a shorter measurement time than pulsed experiments, and a large percentage of all ESR experiments are still carried out with the cw setup shown in Fig. 2a.

- For the RX path, the short relaxation times impose very stringent requirements on the tolerable receiver dead time after the excitation pulse. The required dead times for direct detection of the free induction decay after the excitation pulse (nanosecond range) can today only be met with very significant (i.e., expensive) hardware efforts [10] or using IC-based solutions [11]. Consequently, most pulsed ESR experiments are carried out using so-called echo sequences, which produce an ESR signal with sufficient lag after the pulse.

According to the discussion above, the operating frequencies of typical ESR experiments range from a few gigahertz to approximately 500 GHz, which can mostly be covered by modern IC technologies. Therefore, monolithic integrations of ESR spectrometers have recently gained significant attention in the research community because there is hope that such realizations can remove many of the above mentioned instrumental limitations of existing ESR spectrometers. More specifically, a monolithic integration can drastically reduce the cost of the required electronics and thereby also that of the entire spectrometer. Next, the reduced form factor of a monolithic integration together with the recent advances in magnet technology is a key enabler for portable PoC spectrometers. Finally, a co-integration of the required resonator together with the spectrometer electronics on a single chip can drastically reduce the interconnect distances. This in turn removes the need for impedance matching between PA and resonator as well as resonator and LNA, potentially greatly reducing the required power consumption and opening up entirely new possibilities similar to those discussed for NMR previously.

Despite the many advantages of a monolithic integration of the ESR spectrometer electronics, its major limitation lies in the limited available output power at the elevated ESR frequencies provided by standard IC technologies such as CMOS and bipolar CMOS (BiCMOS). Even when using multiple PAs in combination with passive

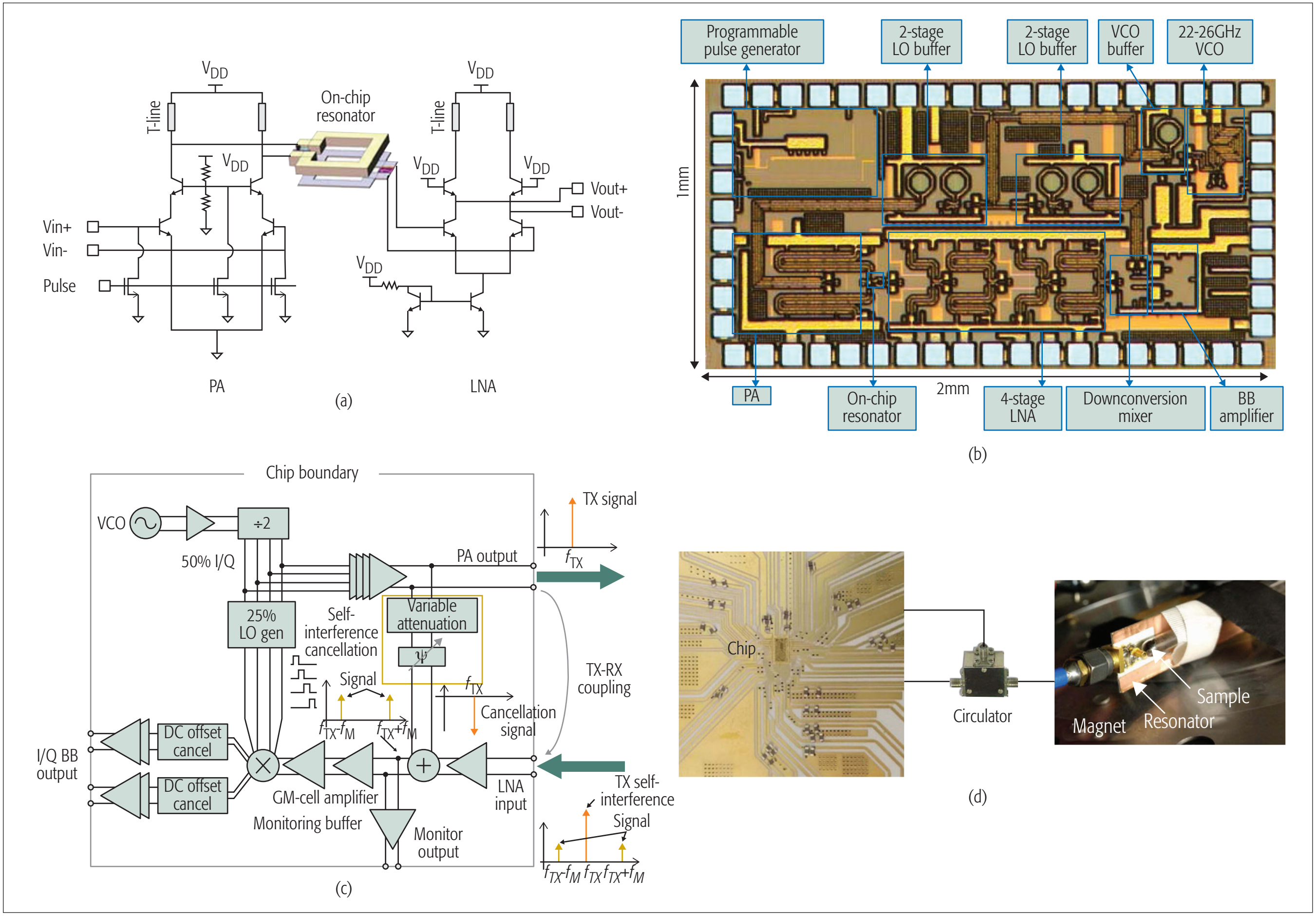


Figure 5. a) Front-end, b) chip micrograph of the single-chip ESR spectrometer presented in [11]; c) architecture, d) photographs of the self-interference cancellation ESR spectrometer presented in [13].

power combining, the available output powers are typically limited to below 30 dBm. While this is certainly sufficient for cw experiments and even pulsed experiments using microresonators [12], it is insufficient for pulsed experiments using larger-volume resonators with volumes greater than approximately 10 μl, as they are required to achieve good concentration sensitivity in the micromolar or even nanomolar range. As a result, as of today, all research activities in the field of IC-based ESR spectrometry are directed to either cw experiments or pulsed experiments using microresonators.

Two examples of these research activities that aim to integrate the components of classical cw and pulsed ESR spectrometers in standard 0.13 μm BiCMOS technology to improve the overall spectrometer cost and form factor are presented in [11, 13] (Fig. 5). The design with a co-integrated microresonator presented in [11] uses separate, concentric, planar resonators for excitation and detection of the ESR signal, and achieves π/2 flip angles with pulse durations of about 100 ns and a dead time as low as 0.5 ns with a resonator with a diameter of 20 μm. However, the very small resonator size limits its use to heavily mass limited samples such as small crystals. In [13], a single-chip ESR spectrometer for use with external, large-volume resonators in cw-ESR experiments was presented. As one highlight of the article, the proposed design features an on-chip self-interference cancellation technique, which improves the isolation between the TX and RX ports.

One of the major limitations of existing benchtop ESR spectrometers is the requirement to sweep the magnetic field to perform a conventional resonator-based cw-ESR experiment, which avoids the intrinsic trade-off between the resonator quality factor and the sweep range in conventional frequency-swept ESR (high Q for high sensitivity, low Q for large sweep range).

To circumvent this problem, a voltage-controlled oscillator (VCO)-based PoC ESR spectrometer was presented by our group in [14]. The novel VCO-based sensing principle extends the oscillator-based detection used in [15] to allow for operation in a fixed field permanent magnet as required by portable PoC applications. The operating principle of VCO-based sensing is illustrated in Fig. 6a. In the steady-state, as it occurs in cw-ESR experiments, the spin ensemble can be modeled as a damped harmonic oscillator represented by the damped LC resonant circuit shown in blue. The interaction between the oscillator and the spin ensemble can be modeled by a mutual inductance with coupling coefficient K_{spin} between the VCO's tank inductance L and the inductance of the LC resonant circuit, L_{spin}, representing the spin ensemble. A VCO-based cw-ESR experiment is performed by sweeping the varactor tuning voltage V_{TUNE} to produce a corresponding change in the VCO's oscillation fre-

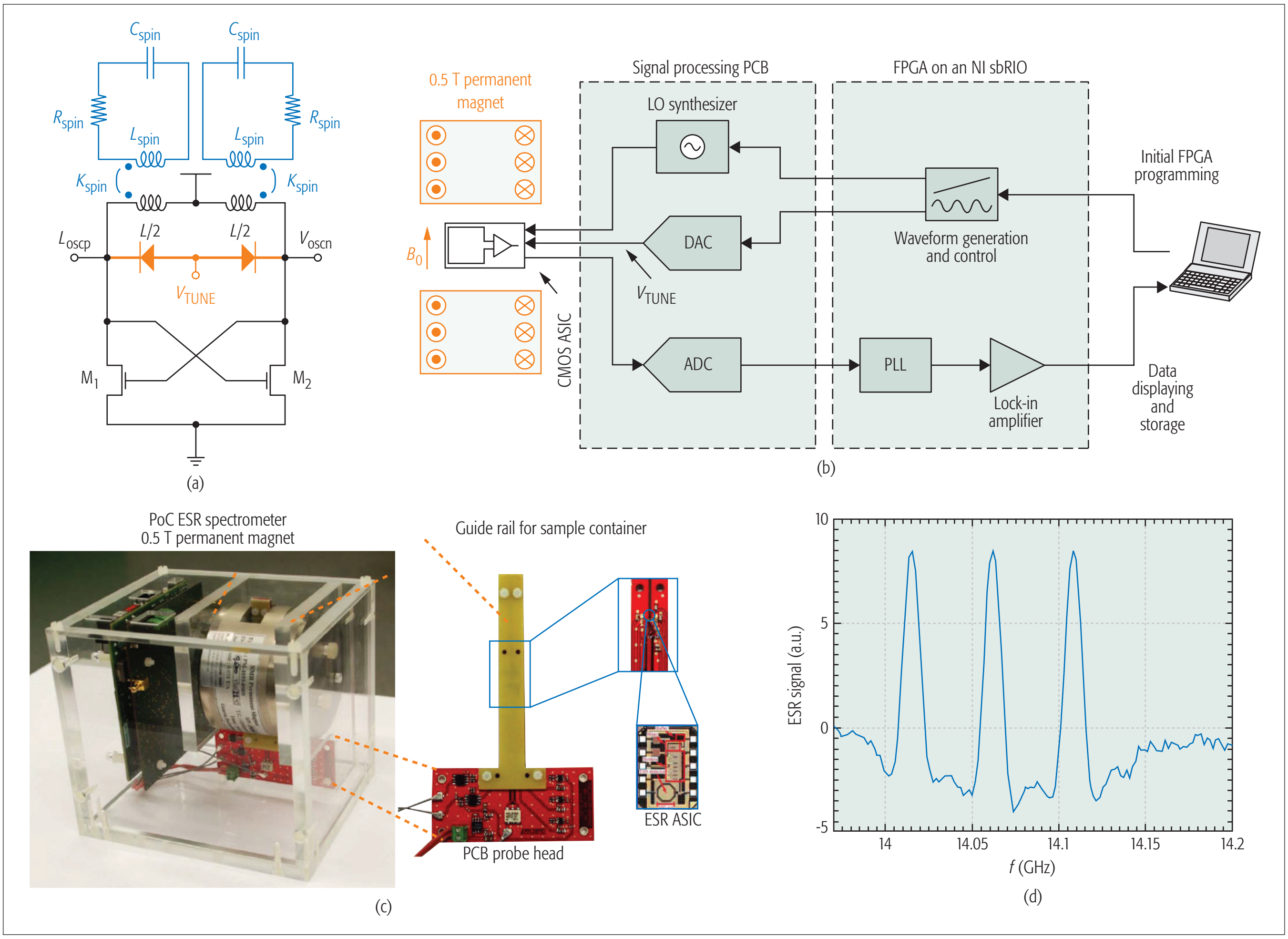


Figure 6. a) Illustration of the VCO-based ESR sensing principle used in [14]; b) spectrometer architecture of the portable point-of-care ESR-spectrometer proposed in [14]; c) photograph of the most recent VCO-based point-of-care ESR spectrometer; d) frequency-sweep ESR spectrum of the spin trap TEMPOL.

quency, that is, $f_{osc} = f_{osc}(V_{TUNE})$. The oscillating current running through the tank inductance L then produces the magnetic field B_1 which resonantly interacts with the spin ensemble if its frequency matches the resonance condition $\omega_{osc} = |\gamma \cdot B_0|$. The major advantage of the VCO-based approach is that a change of the tuning voltage V_{TUNE} changes both the oscillation frequency f_{osc} and the resonance frequency of the LC tank at the same time, causing the LC tank to be critically coupled to the excitation frequency at every point of the frequency sweep, thereby producing a constant detection sensitivity over the entire sweep range. This is in marked contrast to the conventional resonator-based detection where the resonance frequency of the resonator is fixed, and the coupling between the excitation source and the resonator greatly varies as a function of frequency, resulting in very inhomogeneous sensitivity over the sweep range and a sweep range that is limited by the quality factor of the resonator. For the oscillator-based detection, the achievable sweep range is only limited by the tuning range of the utilized VCO. The spectrometer architecture, which at its heart uses an ASIC containing the VCO-based detector described above, is shown in Fig. 6b, and the assembled spectrometer is shown in Fig. 6c. As highlighted in red in Fig. 6b, thanks to the frequency-sweep-based operation, the setup can use a permanent magnet as the source for the 0.5 T static B_0-field, corresponding to a VCO operating frequency around 14 GHz. Moreover, phase-sensitive detection by means of a lock-in amplifier is enabled by frequency instead of field modulation, which removes the need for modulation coils that would require hundreds of milliamperes to produce the required modulation fields. This further reduces the overall power consumption and enables battery operation of the spectrometer. An example ESR spectrum of the spin trap TEMPOL measured using the PoC spectrometer is shown in Fig. 6d. The three peaks in the spectrum result from the coupling of the electron spin with the nuclear spin associated with the nitrogen atom in the TEMPOL molecule. Also in ESR, the integrals over the peaks (or the double integrals when using frequency modulation) are proportional to the number of spins contributing to said peak, rendering ESR a fully quantitative method.

Summary and Conclusion

In this tutorial article, we have discussed the recent trends in the field of general impedance spectroscopy toward the realization of portable PoC spectrometers. We have highlighted that due to the utilized frequency range and the similarity between the architecture of impedance spectroscopy experiments and transceivers for communications applications, the field of impedance spectroscopy can largely benefit from recent improvements in transceiver circuits. Moreover, we have discussed in detail two special cases of impedance spectroscopy, NMR and ESR spectroscopy, which use an external magnetic field to generate very rich spectroscopic information in the RF and microwave frequency range, and therefore ideally lend themselves to IC-based spectrometer realizations. In this context, we have discussed existing IC-based PoC spectrometer realizations for NMR and ESR applications, which illustrate the great potential associated with modern IC technologies for the realization of portable spectrometers. For future generations of PoC spectrometers, the possibilities of IC-based realizations will allow for further system miniaturization, reduced power consumption, and improved sensor performance at reduced spectrometer costs. This in turn will have an enormous impact on the emerging fields of personalized medicine and home diagnostics by making low-cost yet high-performance spectrometers available to a large user community.

For future generations of PoC spectrometers, the possibilities of IC-based realizations will allow for further system miniaturizations, reduced power consumptions and improved sensor performances at reduced spectrometer costs. This in turn will have an enormous impact on the emerging fields of personalized medicine and home diagnostics.

References

[1] F. Vollmer and L. Yang, "Label-Free Detection with High-Q Microcavities: A Review of Biosensing Mechanisms for Integrated Devices," *Nanophotonics*, vol. 1, no. 3–4, 2012, pp. 267–91.

[2] N. Guo *et al.*, "CMOS Time-Resolved, Contact, and Multispectral Fluorescence Imaging for DNA Molecular Diagnostics," *Sensors*, vol. 14, no. 11, 2014, pp. 20,602–19.

[3] M. H. Levitt, *Spin Dynamics: Basics of Nuclear Magnetic Resonance*, 2nd ed., Wiley, 2008.

[4] J. Handwerker *et al.*, "An Array of Fully-Integrated Quadrature TX/RX NMR Field Probes for MRI Trajectory Mapping," *Proc. Euro. Solid-State Circuits Conf.*, 2016, pp. 217–20.

[5] K.-M. Lei *et al.*, "A Handheld High-Sensitivity Micro-NMR CMOS Platform with B-Field Stabilization for Multi-Type Biological/Chemical Assays," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, 2017, pp. 284–97.

[6] M. Grisi, G. Gualco, and G. Boero, "A Broadband Single-Chip Transceiver for Multi-Nuclear NMR Probes," *Rev. Scientific Instruments*, vol. 86, no. 4, 2015, p. 44,703.

[7] J. Kim, B. Hammer, and R. Harjani, "A 5–300 MHz CMOS Transceiver for Multi-Nuclear NMR Spectroscopy," *Proc. IEEE Custom Integrated Circuits Conf.*, 2012, pp. 1–4.

[8] D. Ha *et al.*, "Scalable NMR Spectroscopy with Semiconductor Chips," *Proc. Nat'l. Academy of Sciences*, vol. 111, no. 33, 2014, pp. 11,955–60.

[9] P. P. Stang *et al.*, "Medusa: A Scalable MR Console Using USB," *IEEE Trans. Medical Imaging*, vol. 31, no. 2, 2012, pp. 370–79.

[10] P. A. S. Cruickshank *et al.*, "A Kilowatt Pulsed 94 GHz Electron Paramagnetic Resonance Spectrometer with High Concentration Sensitivity, High Instantaneous Bandwidth, and Low Dead Time," *Rev. Scientific Instruments*, vol. 80, no. 10, 2009, p. 103,102.

[11] C. Chen, P. Seifi, and A. Babakhani, "A Silicon-Based, Fully Integrated Pulse Electron Paramagnetic Resonance System for mm-Wave Spectroscopy," *Proc. IEEE Int'l. Microwave Symp.*, 2013, pp. 1–3.

[12] R. Narkowicz, D. Suter, and R. Stonies, "Planar Microresonators for EPR Experiments," *J. Magnetic Resonance*, vol. 175, no. 2, 2005, pp. 275–84.

[13] X. B. Yang and A. Babakhani, "A Full-Duplex Single-Chip Transceiver with Self-Interference Cancellation in 0.13 m SiGe BiCMOS for Electron Paramagnetic Resonance Spectroscopy," *IEEE J. Solid-State Circuits*, vol. 51, no. 10, 2016, pp. 2408–19.

[14] J. Handwerker *et al.*, "A 14 GHz Battery-Operated Point-of-Care ESR Spectrometer Based on a 0.13 m CMOS ASIC," *Proc. IEEE Int'l. Solid-State Circuits Conf.*, 2016, pp. 476–77.

[15] J. Anders, A. Angerhofer, and G. Boero, "K-Band Single-Chip Electron Spin Resonance Detector," *J. Magnetic Resonance*, vol. 217, 2012, pp. 19–26.

Biographies

Jonas Handwerker [S'10] (jonas.handwerker@uni-ulm.de) received his B.Sc. and M.Sc. degrees in microsystems engineering from IMTEK, University of Freiburg, Germany, in 2008 and 2011, respectively. From 2009 to 2010, he interned at the Robert Bosch Research and Technology Center, Palo Alto, California. Since 2012, he has been working toward a Ph.D. degree at the University of Ulm, Germany. His research interests include IC and MEMS design for NMR imaging and spectroscopy, and MRI gradient field correction.

Benedikt Schlecker [S'12] (benedikt.schlecker@uni-ulm.de) received his Dipl.-Ing. degree with honors in electrical engineering from the University of Ulm in 2012. For his outstanding Master's thesis he received the VDE-Förderpreis 2012. Currently, he is working toward a Ph.D. degree at the University of Ulm in the field of readout and demodulation circuits for low-noise, high-speed sensing applications.

Maurits Ortmanns [SM'11] (maurits.ortmanns@uni-ulm.de) received his Dr.-Ing. from the University of Freiburg in 2004. From 2006 to 2007 he was an assistant professor of integrated interface circuits at the University of Freiburg, and since 2008 he has been director of the Institute of Microelectronics at the University of Ulm. His research interests include circuit design for data converters and biomedical applications. He holds several patents, and has published several books and more than 200 IEEE journal and conference papers.

Jens Anders [SM'17] (jens.anders@uni-ulm.de) received his Ph.D. from EPFL Lausanne in 2011. Since 2013 he has been an assistant professor of biomedical integrated sensors within the Institute of Microelectronics at the University of Ulm. He is the recipient of some nationwide scientific awards in Germany and the author/coauthor of several books and book chapters as well as approximately 100 journal and conference papers. His research interests include circuit design for materials science and biomedical sensing.

Multi-Service System: An Enabler of Flexible 5G Air Interface

Lei Zhang, Ayesha Ijaz, Pei Xiao, and Rahim Tafazolli

The authors present a framework for a multi-service system, and the challenges and possible solutions are studied. The multi-service system implementation in both the time and frequency domains is discussed. Two representative SFMC waveforms, F-OFDM and UFMC, are considered. Specifically, the design methodology, criteria, orthogonality conditions, and prospective application scenarios in the context of 5G are discussed.

Abstract

A multi-service system is an enabler to flexibly support diverse communication requirements for the next generation wireless communications. In such a system, multiple types of services coexist in one baseband system with each service having its optimal frame structure and low out-of-band emission waveforms operating on the service frequency band to reduce the ISvcBI. In this article, a framework for a multi-service system is established, and the challenges and possible solutions are studied. The multi-service system implementation in both the time and frequency domains is discussed. Two representative SFMC waveforms, F-OFDM and UFMC, are considered in this article. Specifically, the design methodology, criteria, orthogonality conditions, and prospective application scenarios in the context of 5G are discussed. We consider both SR and MR signal processing methods. Compared to the SR system, the MR system has significantly reduced computational complexity at the expense of performance loss due to ISubBI in MR systems. The ISvcBI and ISubBI in MR systems are investigated with proposed low-complexity interference cancellation algorithms to enable multi-service operation in low interference level conditions.

Introduction

Fifth generation (5G) wireless communication systems are expected to address unprecedented challenges to cope with a high degree of heterogeneity in terms of services, device classes, deployment environments, and mobility levels [1]. Different applications and use cases specified by the 5G research community have been categorized into three main communication scenarios [2]: enhanced mobile broadband (eMBB), massive machine type communications (mMTC), and ultra-reliable and low latency communications (URLLC).

Designing a separate standalone radio system for each service to support heterogeneous requirements is not a feasible solution, since the operation and management of the systems will be highly complex, expensive, and spirally inefficient. On the other hand, it is cumbersome to design a unified all-in-one radio frame structure that meets the requirements for all types of services. For example, mMTC may require smaller subcarrier spacing (and thus larger symbol duration) to support massive delay-tolerant devices. URLLC, on the other hand, has more stringent reliability and latency requirements; thus, symbol duration must be significantly reduced. However, the subcarrier spacing and symbol duration of eMBB communication are constrained by a doubly dispersive channel (i.e., channel coherence time and coherence bandwidth). Therefore, there is a limit on subcarrier spacing and symbol duration in order to avoid performance bottlenecks due to channel impairments.

One viable solution to support diverse requirements in 5G is to multiplex the multiple types of services in one baseband system in orthogonal time and/or frequency resources, with either physical (e.g., using a guard interval or guard band) or algorithmic (e.g., filtering or precoding the data) isolation to avoid the interference between them [3, 4]. Frequency-division multiplexing (FDM) is preferred by the Third Generation Partnership Project (3GPP) for multiplexing different services due to several advantages including good forward compatibility, ease of supporting services with different latency requirements, energy saving by turning off some transmit time intervals (TTIs), and so on. Such an FDM multi-service system is shown in Fig. 1a, where an optimal frame structure has been designed for different types of services in different service frequency bands, with a low out-of-band emission (OoBE) subband filtering operation to reduce the interference. An optional guard band could be used between them to further mitigate the interference.

In addition to economic benefits and dynamic resource allocation, the multi-service approach exclusively optimizes the parameters to cater for the unique service requirements in each scenario. Moreover, multi-service systems can enable loose time synchronization and may save signaling overhead, for example, time advance (TA) in Long Term Evolution (LTE), since all service signals are well separated in the frequency domain. The spectrum allocation flexibility of the multi-service system can also be combined with other techniques such as cognitive radio networks [5–7], where the fragmented spectrum can be dynamically occupied by various types of services and keep the services from significant inter-service-band interference (ISvcBI).

It can be verified from mathematical analysis that combining different numerologies in one frequency band will destroy the orthogonality of multi-carrier systems, resulting in ISvcBI. Inserting a guard band between service bands can miti-

Digital Object Identifier:
10.1109/MCOM.2017.1600916

Lei Zhang is with the University of Glasgow; Ayesha Ijaz, Pei Xiao, and Rahim Tafazolli are with the University of Surrey.

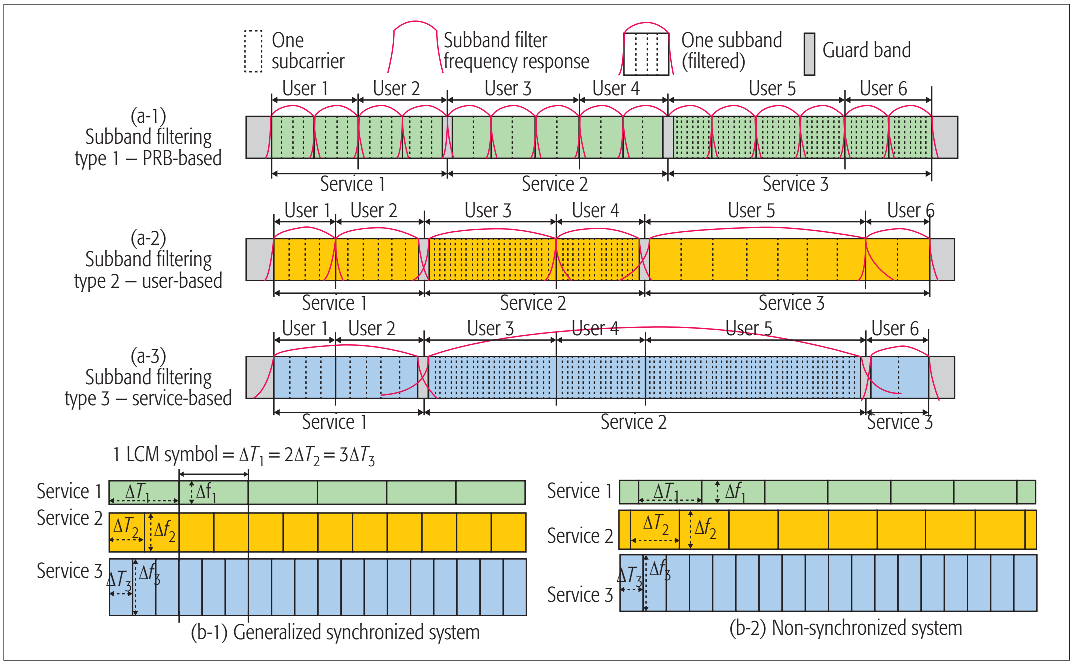


Figure 1. Multi-service system frequency and time domain implementations: a) three types of subband filtering methods; b) generalized synchronized and non-synchronized multi-service systems.

The multi-service system may fundamentally change the air interface architecture and algorithms employed in existing single-service systems. These changes and extensions may require rethinking the availability and effectiveness of using existing design criteria, algorithms, optimization, and performance analysis for multi-service systems.

gate the interference, but at the cost of reduced radio spectrum efficiency. Waveforms with low OoBE are important in the multi-service system in order to isolate the signals between services and reduce the ISvcBI with/without a limited guard band between them. Several new waveforms have been proposed for next generation communications with OoBE level as the most important key performance indicator (KPI). Among them, filtered orthogonal frequency-division multiplexing (F-OFDM) [4] and universal filtered multi-carrier (UFMC) [3, 8] are particularly promising due to their excellent trade-off between complexity and performance. Thus, they were investigated as the main candidate waveforms for 5G in the 3GPP RAN1 meeting [9].

The multi-service system may fundamentally change the air interface architecture and algorithms employed in existing single-service systems (e.g., OFDM-based LTE). These changes and extensions may require rethinking the availability and effectiveness of using existing design criteria, algorithms, optimization, and performance analysis for multi-service systems. Specifically, the multi-service system is different in the following aspects:

- Even with low OoBE waveforms, the multi-service system is no longer orthogonal due to the trade-off between the performance and system overhead. Inter-symbol interference (ISI) and ISvcBI exist in the system.
- Due to the subband filtering, the filter gain at different subcarriers in one subband may be different, resulting in uneven power allocation among subcarriers and hence performance loss [3].
- Multi-rate (MR) implementation may be essential to make the multi-service system complexity affordable [10]. However, compared to single-rate (SR) implementation, MR may degrade the system performance due to the inter-subband-interference (ISubBI) generated in the up/down-sampling process.
- F-OFDM and UFMC are designed by maximizing the frequency and time localization property, respectively, resulting in the two waveforms favoring different application scenarios.

All of the aforementioned aspects will be systematically discussed in this article to provide guidelines for the 5G system design and solutions to network slicing on physical layer resource multiplexing and isolation. Note that this article focuses on the fundamental limitations and applicable scenarios for multi-service systems based on F-OFDM and UFMC waveforms. The original waveform signal model can be found in [4, 10], while the mathematical model of a multi-service system and the details of algorithms used in the article can be found in [3, 10]. It must be noted that in a single-service system such as LTE with single numerology, inter-carrier-interference (ICI) defines the interference generated among the subcarriers. However, ICI is not sufficient to capture all the impairments incurred in a multi-service system, where different services may use different subcarrier spacing and symbol duration. The ICI definition, analysis, and cancelation algorithms in the traditional single-service system cannot be applied to the multi-service system. To differentiate it, we define the *interference between service bands as ISvcBI and the interference between subbands in one service band as ISubBI.*

Note that [11] proposed a multi-service system called flexible configured OFDM (FC-OFDM) by using time domain windowing to reduce the system OoBE and a novel low-complex precoding (with 2 taps only) to mitigate the interference. However, it may result in higher ISvcBI, and a large guard band may be required to reduce the interference level in edge subcarriers. In addition, [12] proposed a multi-service system based on

Besides the complexity and flexibility considerations, granularity of the subband also depends on the employed waveform. Waveforms with better frequency but worse time localization property may favor user- or service-based implementation. On the other hand, a waveform with better time but worse frequency localization may prefer the PRB-based implementation.

the filter-bank multicarrier (FBMC) waveform that may provide a better OoBE and isolation between service bands. However, as also pointed out in the literature [3, 4, 11, 13], the FBMC system is significantly more complex than an OFDM-based system. Nevertheless, the proposed interference cancellation schemes are generic and can be combined with other systems such as FC-OFDM and FBMC proposed in [11, 12], respectively.

In this article, we build a framework for a multi-service system and categorize the possible subband filtering implementations and synchronized systems in the frequency and time domains. The roles of the waveform and subband filter in the multi-service system are discussed, and the two waveforms' limitations and viable subband bandwidth regions are also discussed. The waveforms' prospective application scenarios in the context of 5G are investigated. We also discuss single-rate and multi-rate implementations of multi-service system. The system orthogonality and the sources of the ISvcBI and ISubBI are discussed in detail. In addition, the ISvcBI and ISubBI cancellation algorithms and simulation results are presented.

In this article, we use the following parameters for numerical evaluations unless otherwise specified:

- 20 MHz system bandwidth and 30.72 MHz sampling rate contain 2048 subcarriers.
- Zero padding (ZP) or cyclic prefix (CP) length is 160 samples.
- The respective filter for F-OFDM and UFMC is the Windowed Sinc filter [4] and Chebyshev filter (with OoBE being –50 dB) [13], and the filter length is 1024 and 160 samples, respectively
- We consider the International Telecommunication Union (ITU)-defined urban micro (UMi) channel for all simulations.

Multi-Service System Implementations

Multi-Service System Frequency Domain Implementation

For a multi-service system, it is natural to assume that each service supports one or more users, where each user can be granted an arbitrary number of consecutive or non-consecutive physical resource blocks (PRBs). The possible bandwidth allocation and subband filtering methods in a multi-service system are shown in Fig. 1a. The conventional multi-carrier systems (e.g., LTE/LTE-Advanced, LTE-A) have a three-tier frequency resource structure, that is, system bandwidth, PRB, and subcarrier. However, the multi-service system has a four-tier frequency resource structure, that is, system bandwidth, service bandwidth, PRB, and subcarrier. The level on which the subband filter operates will affect the multi-service system performance and implementation complexity. Figures 1a-1, 1a-2, and 1a-3 show filtering applied to PRB, user, and service, respectively.

Each subband filtering scheme has its own pros and cons. The PRB is the minimum scheduling granularity, and the subband filtering based on one or more PRBs (Fig. 1a-1) has maximum design flexibility. On the other hand, this implementation also incurs the highest computational complexity due to the dense subband filtering operation. On the contrary, the service-based subband filtering method (Fig. 1a-3) has the lowest computational complexity, and the users (and PRBs) in one service share the same filter design parameters. Hence, the system loses the advantage of independently optimized filter design to cater for the specific scenarios. User-based subband filtering as shown in Fig. 1a-2 is a trade-off between PRB-based and service-based methods. Note that PRB-based implementation is the most general case.

Besides the complexity and flexibility considerations, granularity of the subband also depends on the employed waveform. Waveforms with better frequency but worse time localization property (e.g., F-OFDM) may favor user- or service-based implementation. On the other hand, a waveform with better time but worse frequency localization (e.g., UFMC) may prefer the PRB-based implementation. This is discussed next in detail.

Multi-Service System Time Domain Implementation

Since the symbol duration is different for different services, this makes the (spectrally efficient) synchronization of the whole system practically impossible. For example, in OFDM systems, without considering the guard interval, having two services with subcarrier spacing $\Delta f_2 = 2\Delta f_1$ implies that the symbol duration has the relationship $\Delta T_1 = 2\Delta T_2$ (Fig. 1b-1). Consequently, the symbols in service 2 cannot synchronize with symbols in service 1. However, we can take advantage of the fact that the duration of every 2 symbols in service 2 is the same as the symbol duration in service 1; we call this a generalized synchronized (GS) system. In such a system, there is a duration equivalent to the least common multiple (LCM) of symbol durations of all services. Figure 1b-1 is an example of the GS system, which has the advantage of simplified system and algorithm design and performance analysis since only limited symbols need to be considered in a processing window, and every LCM window has the same overall performance.

However, in a GS system, the symbol duration plus overhead (filter tails, guard interval, etc.) for all services should have an LCM, which might reduce the system design flexibility. Moreover, all services have to be synchronized to take advantage of the GS system. Therefore, a non-synchronized MS system as given in Fig. 1b-2 may be considered in some scenarios.

Waveform Design and Comparisons

F-OFDM and UFMC Design Criteria

According to the Balian-Low Theorem [14], there is no way to utilize a well localized prototype filter in both time and frequency, along with maintaining orthogonality and transmitting at the Nyquist rate. Hence, relaxing one condition guarantees the other two factors. UFMC and F-OFDM are two contrasting examples. The former uses a short filter to secure a good time localization property. In such a case, the ISI can be minimized, but the sacrificed filter frequency localization property may generate more ISvcBI/ISubBI in multi-service systems. While F-OFDM uses a long filter with sharp cut-off resulting in ISvcBI/ISubBI minimization, this may generate ISI, which could be significant in some scenarios such as narrowband mMTC communications.

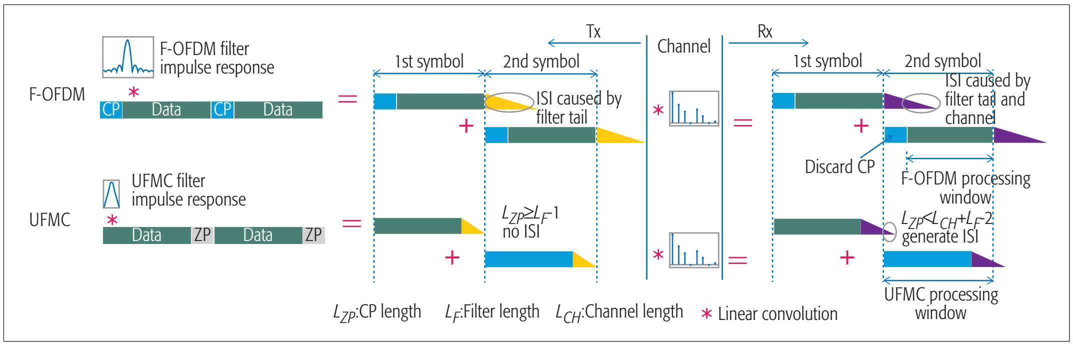


Figure 2. F-OFDM and UFMC system implementations.

Compared to OFDM systems, SFMC systems may suffer from filter frequency response selectivity among subcarriers. This side-effect causes power allocation imbalance and performance loss if all subcarriers carry equally important information. This effect may be especially detrimental for the UFMC system.

The time domain implementations of both F-OFDM and UFMC are shown in Fig. 2, where only one subband and two consecutive symbols are considered for demonstration purposes. Essentially, the UFMC is a ZP-based multi-carrier system and the F-OFDM is the CP-based one. The UFMC symbols do not overlap at the transmitter. However, this does not mean that UFMC is an ISI-free system since the adjacent symbols will overlap after passing through a multipath channel as shown in Fig. 2. In F-OFDM systems, a longer filter is used and filter tails extend to adjacent symbols [4]. Overlapping and ISI are unavoidable for a reasonable system overhead. At the receiver side, the UFMC and F-OFDM can use the standard ZP- or CP-based multicarrier system processing with a matched filter as an option.

Filter Length, CP/ZP Length Selection, and Impact on ISI/ICI

CP/ZP plays an important role in the OFDM system in terms of spectrum efficiency and performance. It can eliminate the ISI and allows low-complexity interference-free one-tap channel equalization if only the guard interval is equal to or longer than the channel length. This condition, however, is not sufficient for F-OFDM and UFMC systems.

State-of-the-art (SoTA) UFMC constrains the ZP length and the filter length to be equal to the channel length to trade off the system overhead and performance [8, 12]. In such a case, the reserved ZP at the transmitter will be occupied by the filter tail completely. Although the filter ramp-up and ramp-down may mitigate the multipath channel effects to some degree, it cannot eliminate ISI completely.

In fact, the ZP length and the filter length can be de-coupled to optimize the system performance. For example, with a fixed overall system overhead, one can design a system with smaller filter length (thus, short filter tail) and leave some degree of freedom (i.e., zero at the end of the symbol) to mitigate the multipath channel dispersion. This might be especially useful for the symbol with pilot subcarriers for channel estimation.

With the short filter length and good time localization property, the UFMC system may suffer from more ISvcBI/ISubBI and performance loss due to inefficient power allocation in the multi-service system, which is shown later in this article.

The CP length in F-OFDM is normally set to be the same as the channel length. However, the filter length can be as long as half symbol duration [4]. This design criterion provides very good frequency localization in the F-OFDM system. Allowing adjacent symbols to overlap at the transmitter side might subject the F-OFDM system to ISI contamination. However, filter impulse response decays significantly. In addition, the CP absorbs most of the energy of the filter if the subband bandwidth is not extremely small [4]. However, F-OFDM may require longer CP in narrowband systems to mitigate the ISI. Figure 3a shows the ISI vs. the normalized subband bandwidth for different CP lengths (LCP) in the F-OFDM system in the ITU UMi channel. It can be seen that a larger subband bandwidth leads to less ISI, and an increase in the CP length can significantly reduce the interference level.

Waveform Filter Frequency Selectivity and Impact on Performance

Compared to OFDM systems, SFMC systems may suffer from filter frequency response selectivity among subcarriers. This side-effect causes power allocation imbalance and performance loss if all subcarriers carry equally important information. This effect may be especially detrimental for the UFMC system [3]. In particular, the passband bandwidth of the subband filter (e.g., Chebyshev filter) cannot be dynamically changed over a large range due to the short filter length, resulting in limited flexibility in the UFMC system design.

Figure 3b shows the relationship of the filter length with the subband bandwidth for different filter peak-to-bottom-gain ratio (PBGR) (i.e., the ratio of the maximum and minimum filter gain among all subcarriers within one subband) [3]. Note that PBGR = 0 dB means there is no frequency selectivity among the subbands. In this case, UFMC degrades to an OFDM system. Figure 3b shows that a longer filter results in a larger PBGR and greater performance loss. In addition, narrower subband bandwidth results in a smaller PBGR and thus better performance. Figure 3b can be used in multiple ways for the design of UFMC-based 5G systems. For example, we can select appropriate subband bandwidth to achieve a certain PBGR for a given total number of subcarriers and filter length. Similarly, for given filter length and subband bandwidth, it is easy to calculate corresponding PBGR, based on which the performance loss can be evaluated.

The frequency selectivity may also affect the channel estimation algorithms and optimal pilot pattern design. It is preferable to assign pilots at the subcarriers with the largest filter gain (i.e., in the middle of one subband). In addition, a tradi-

In the SR system, the orthogonality between the subcarriers in one service is ensured by taking the corresponding columns of the full-size inverse discrete Fourier transform modulation. One of the important roles of a subband filter is to reduce the ISvcBI among services. Such a system may have very high computational complexity.

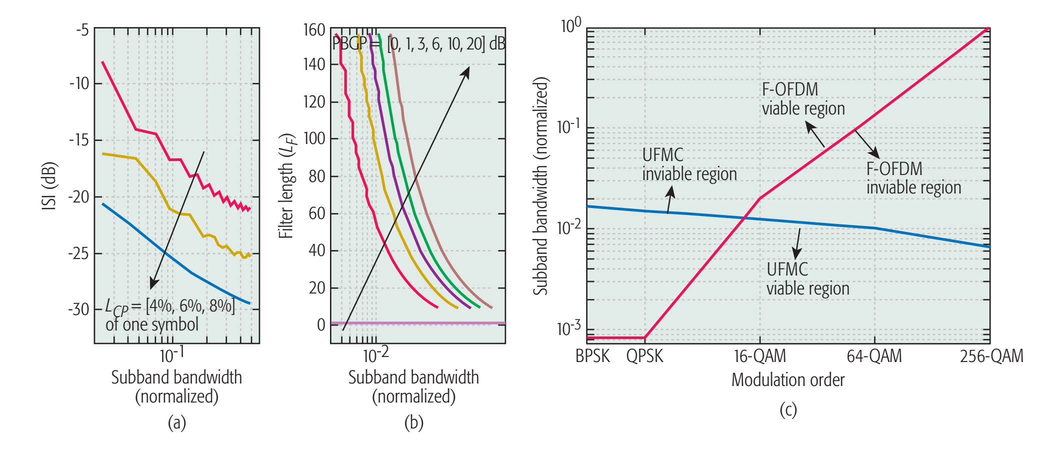


Figure 3. F-OFDM and UFMC performance in terms of subband bandwidth: a) ISI vs. subband bandwidth with different CP length for F-OFDM; b) filter length vs. subband bandwidth with different PBGR for UFMC; c) viable (subband bandwidth) region of F-OFDM and UFMC.

tional channel estimation algorithm such as polynomial interpolation is no longer suitable for the SFMC system.

Waveforms' Viable Subband Bandwidth Regions

According to the earlier discussion, an F-OFDM system is a subband bandwidth low-bounded system and UFMC is a subband bandwidth high-bounded system. Figure 3c shows simulation results illustrating the bounds and the viable subband bandwidth region of the two waveforms in the ITU UMi channel for different modulation levels in order to reach 10^{-3} or lower uncoded bit error rate (BER). It can be seen that when modulation levels are low, both waveforms have larger viable ranges. As the modulation level increases, the viable subband bandwidth tends to decrease. With given ZP/CP length and system bandwidth, Fig. 3c implies that small subband bandwidth is a more suitable region for UFMC since it has smaller filter gain frequency selectivity and thus smaller overall performance loss. F-OFDM, on the other hand, prefers to use larger subband bandwidth to protect the system from ISI contamination.

The viable region directly relates to the design flexibility and complexity. Small subband bandwidth may bring more degrees of freedom in the design (e.g., narrowband mMTC services). For this reason, the F-OFDM system may have limited applications. For example, F-OFDM can only support a single service with 256-quadrature amplitude modulation (QAM) to achieve the target BER, whereas up to 100 different subbands/services can be supported in UFMC. However, too small subband bandwidth leads to higher computational complexity. In addition, in the eMBB/ URLLC scenario, a relatively larger subband may be granted to one user. Thus, multiple subbands for one user may lead to unnecessary complexity. In such a scenario, F-OFDM is the preferred choice.

SR and MR Implantation of Multi-Service Systems

There are two implementations for the multi-service SFMC system: SR and MR. Compared to an SR system, an MR system has significantly reduced computational complexity but may suffer loss in performance due to ISubBI. Implementations and comparisons are studied next with a conclusion on their prospective application scenarios.

SR and MR System Orthogonality Analysis

In the SR system, as shown in Figs. 4a and 4b, the orthogonality between the subcarriers in one service is ensured by taking the corresponding columns of the full-size inverse discrete Fourier transform (IDFT) modulation [10]. One of the important roles of a subband filter is to reduce the ISvcBI among services. Such a system may have very high computational complexity.

Alternatively, an MR system reduces the system complexity by up- and down-sampling the signals. As shown in Figs. 4c and 4d, it uses low-dimension full-size IDFT (DFT size is the same as the number of subcarriers in one subband, e.g., 12) that spreads the signal into the whole baseband bandwidth. The following up-sampling operation squeezes the signal into $1/Q_i$ of the full bandwidth with $(Q_i – 1)$ image signals in adjacent bands. An anti-image subband filter is required to mitigate the image signals (i.e., ISubBI) [10]. Nevertheless, the residual image signal will create the ISubBI in the system due to non-ideal filters, which may degrade system performance in comparison to the SR.

Note that the ISubBI is generated on both the transmitter and receiver sides if both use the MR implementations. However, one can use the MR implementation on one side and SR on the other. For example, by using the computational capability advantage at the base station, we can implement the SR at the base station and MR at the mobile station. In addition, we can build a hybrid system by using SR in some subbands with high communication qaulity of service (QoS) requirements (e.g., eMBB) and MR implementations in others that require low computational complexity (e.g., mMTC).

Computational Complexity of the SR and MR Systems

The transmitter computational complexity in terms of the real multiplication of the MR and SR systems for both waveforms is shown in Fig. 5 (the detailed calculation methods can be found in [3, 10]). Note that the complexity is based on one

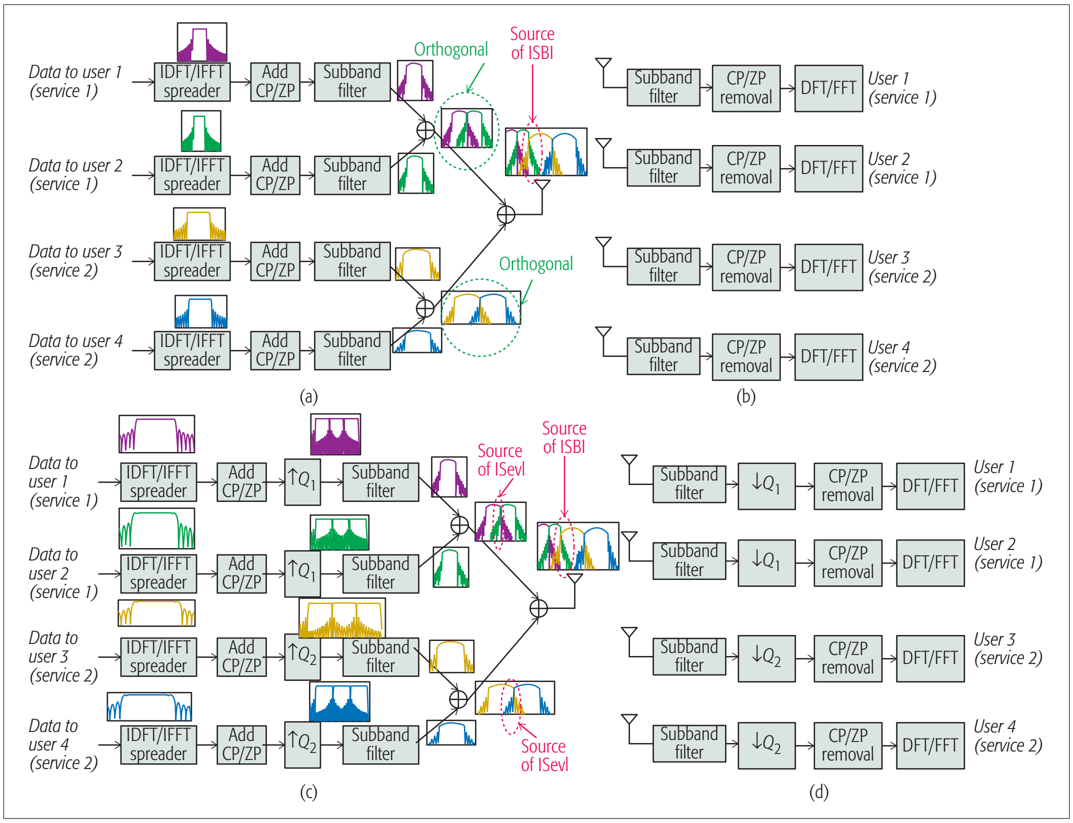


Figure 4. Transmitter and receiver block diagram of SR and MR multi-service systems. For brevity, we consider 4 users in this diagram. Users 1 and 2 belong to service 1, and users 3 and 4 belong to service 2: a) SR multi-service SFMC transmitter; b) SR multi-service SFMC receivers; c) MR multi-service SFMC transmitter; d) MR multi-service SFMC receivers.

service, and it is normalized by the complexity of the OFDM system. The subband bandwidth for UFMC is 16 subcarriers, and there is only one subband in F-OFDM (i.e., it is a service-based implementation as shown in Fig. 1a-3).

The subband filtering can be implemented by either following the traditional linear convolution in the time domain (TD), or by using fast Fourier transform (FFT) in the frequency domain (FD). In MR, the TD subband filtering can take the computational complexity advantage of up-sampling operation since the data is sparse [10]. For the UFMC system, we can see that SR implementation complexity is significantly (up to 1000 times) higher than OFDM system, while the MR system with TD filtering can achieve comparable complexity to the OFDM system. On the other hand, the complexity reduction in F-OFDM by using MR implementation is less significant in a large service band region since there is only one subband in the service. FD filtering is essential for both SR and MR implementations due to the long filter setup in an F-OFDM system.

ISvcBI and ISubBI Cancellation Algorithms for Multi-service Systems

Using a guard band between service bands/subbands can mitigate the ISvcBI/ISubBI, but at the expense of spectrum efficiency reduction. In the following, we propose the baseband signal processing method to cancel ISvcBI/ISubBI on either the transmitter or receiver side.

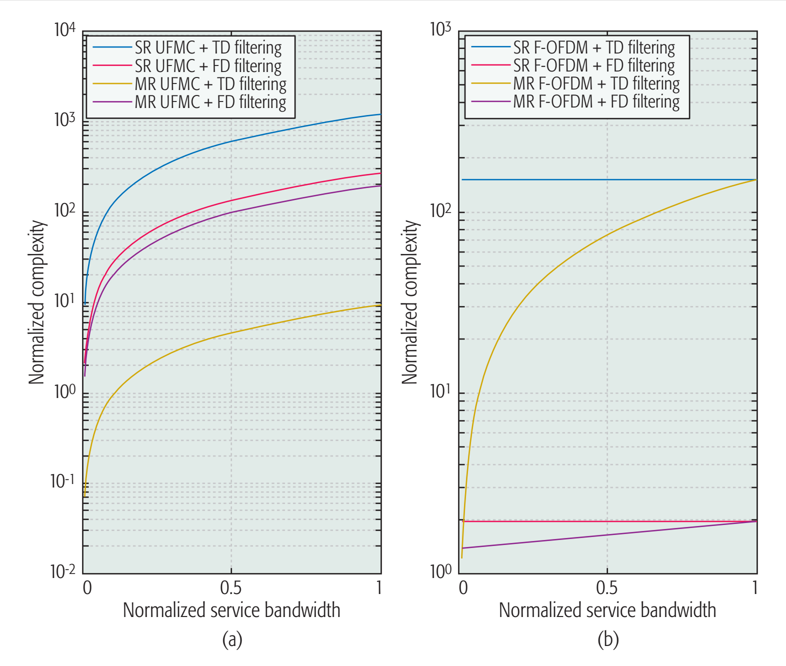


Figure 5. UFMC (left) and F-OFDM-based (right) multi-service system computational complexity (normalized by the OFDM system).

ISvcBI Cancellation Algorithms

Usually, the information carried in two service bands belongs to two different users. Thus, it is difficult to cancel the interference on the user side. In addition, the BS has much higher computation-

Similar to ISvcBI, non-adjacent subbands do not generate significant ISubBI and affect performance. Therefore, we only consider subbands adjacent to each other in the frequency band. In addition, we can use a low-complexity channel-independent ISubBI cancellation algorithm.

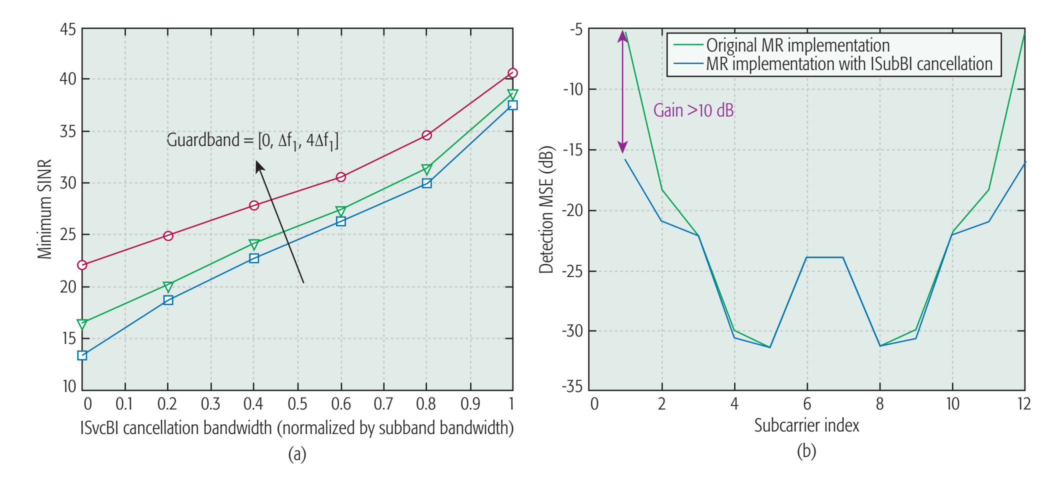


Figure 6. Multi-service system performance with ISvcBI and ISubBI cancellation (each subband contains 12 subcarriers): a) minimum SINR among subcarriers in subbands vs. ISvcBI cancellation bandwidth; b) MSE vs. subcarrier index for original and proposed MR with ISubBI cancellation.

al capability to deal with interference. Therefore, pre-processing the transmit signal at the transmitter in downlink or joint detection in uplink at the receiver can be proposed to cancel this type of interference.

Note that non-adjacent service bands do not generate significant ISvcBI or affect the performance. For example, in Fig. 1a-1, the fourth and fifth subbands located at the edge of the first and second services may generate and suffer from severe ISvcBI. However, the third subband does not generate ISvcBI in the fourth subband, which acts as a buffer zone attenuating the interference from subband 3 to subband 5. In addition, for the fourth and fifth subbands, due to the fast attenuation of the filter response in the stopband, only some subcarriers (e.g., the last subcarrier of the fourth subband and the first subcarrier of the fifth subband in Fig. 1a-1) at the edge of service bands may suffer from severe interference.

The optimal interference cancellation solution should be channel-dependent. Fortunately, the considered bandwidths contaminated by ISvcBI are adjacent to each other, and the interference level decreases exponentially in the subcarriers away from the edge of the service band. Therefore, the channel response for all subcarriers, considered for ISvcBI cancellation, is approximately the same, resulting in a simplified algorithm that does not depend on the channel [3]. Therefore, the solution can be calculated offline in advance to save the computational complexity. For the detailed ISvcBI cancellation algorithms, please refer to [3].

The minimum signal-to-interference-plus-noise ratio (SINR) (worst case) among the subcarriers in one subband (i.e., the edge subcarrier in the edge subband of one service band) vs. the processing bandwidth (normalized by the subband bandwidth) is shown in Fig. 6a for different values of guard band. The results are based on UFMC, and we set the input signal-to-noise ratio (SNR) = 50 dB to limit the system interference. The two considered subbands' subcarrier spacing has the relationship, and each subband has 12 subcarriers. Note that processing bandwidth being zero means no ISvcBI cancellation algorithm is used in the system. Figure 6a shows that larger GB leads to better output SINR. With the ISvcBI cancellation algorithm, the performance can be significantly improved.

ISubBI Cancelation Algorithms

Similar to ISvcBI, non-adjacent subbands do not generate significant ISubBI and affect performance. Therefore, we only consider subbands adjacent to each other in the frequency band. In addition, we can use a low-complexity channel-independent ISubBI cancellation algorithm [10]. Figure 6b shows the proposed ISubBI cancelation algorithm for UFMC performed at the transmitter by precoding the transmit signals, where only two subcarriers at the edge are considered for the ISubBI cancellation as an example. One can see from the figure that the system performance after interference cancellation shows significant gain compared to the one without interference cancelation.

Conclusions and Future Works

A framework for multi-service system is established based on subband filtered multicarrier modulation. The subband filtering implementations of the multi-service system have been discussed. The waveforms' design criteria, orthogonality, and fundamental limitation are studied with the conclusion that filtered orthogonal frequency-division multiplexing may favor user- or service-based subband filtering for enhanced mobile broadband/ ultra-reliable and low latency communications. Universal filtered multicarrier is suitable for physical-resource-block-based subband filtering and massive machine type communications. We consider both single-rate and multi-rate signal processing with detailed analysis of inter-service-band interference and inter-subband interference. The proposed low-complexity ISvcBI and ISubBI cancellation algorithm can significantly improve the system performance with a limited guard band between subbands.

The future work on multi-service system includes, but is not limited to, the following topics:

- Design of new optimal channel estimation and equalization algorithms for the multi-service system by taking the waveform filter frequency selectivity into account

•• Low-complexity interference cancellation algorithms for multiple-input multiple-output cases
- Proposals of new synchronization algorithms in the presence of the non-orthogonal waveforms in multi-service systems
- Mixed/hybrid MR and SR systems, and/or mixed waveforms among service bands

In addition, network slicing has been proposed recently in order to maximize the network utilization and reduce the operational expenditure [15]. The work presented in this article shows how the network slicing can be underpinned in the physical layer in terms of signal multiplexing and isolation. Further technical challenges and potential applications of physical layer network slicing could be a research topic in the future as well.

The work presented in this article shows how the network slicing can be underpinned in the physical layer in terms of signal multiplexing and isolation. Further technical challenges and potential applications of physical layer network slicing could be a research topic in the future as well.

Acknowledgment

The authors would like to acknowledge the support of the University of Surrey 5GIC (http://www.surrey.ac.uk/5gic) members for this work.

References

[1] F. Schaich *et al.*, "FANTASTIC-5G: 5G-PPP Project on 5G Air Interface below 6 GHz," *Proc. Euro. Conf. Network Commun.*, June 2015.

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

[3] L. Zhang *et al.*, "Subband Filtered Multi-Carrier Systems for Multi-Service Wireless Communications," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, Mar. 2017, pp. 1893–1907.

[4] X. Zhang *et al.*, "Filtered-OFDM — Enabler for Flexible Waveform in the 5th Generation Cellular Networks," *IEEE GLOBECOM*, 2015, pp. 1–6.

[5] G. A. Shah, V. C. Gungor, and O. B. Akan, "A Cross-Layer QoS-Aware Communication Framework in Cognitive Radio Sensor Networks for Smart Grid Applications," *IEEE Trans. Industrial Informatics*, vol. 9, no. 3, Aug. 2013, pp. 1477–85.

[6] S. H. R. Bukhari, M. H. Rehmani, and S. Siraj, "A Survey of Channel Bonding for Wireless Networks and Guidelines of Channel Bonding for Futuristic Cognitive Radio Sensor Networks," *IEEE Commun. Surveys & Tutorials*, vol. 18, no. 2, 2016, pp. 924–48.

[7] F. Akhtar, M. H. Rehmani, and M. Reisslein, "White Space: Definitional Perspectives and Their Role in Exploiting Spectrum Opportunities," *Telecommun. Policy*, vol. 40, no. 4, 2016, pp. 319–31.

[8] 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.

[9] 3GPP RP-160671, "New SID Proposal: Study on New Radio Access Technology," TSG RAN Metting 71, Mar. 2016.

[10] L. Zhang *et al.*, "Single-Rate and Multi-Rate Multi-Service Systems for Next Generation and Beyond Communications," *IEEE PIMRC*, Sept. 2016, pp. 1–6.

[11] H. Lin, "Flexible Configured OFDM for 5G Air Interface," *IEEE Access*, vol. 3, , 2015, pp. 1861–70.

[12] M. Fuhrwerk, J. Peissig, and M. Schellmann, "On the Design of an FBMC Based AIR Interface Enabling Channel Adaptive Pulse Shaping per Sub-Band," *Proc. Euro. Signal Processing Conf.*, Nice, France, 2015, pp. 384–88.

[13] V. Vakilian *et al.*, "Universal-Filtered Multi-Carrier Technique for Wireless Systems Beyond LTE," *Proc. IEEE GLOBECOM Wksps.*, 2013, pp. 223–28.

[14] A. Sahin, I. Guvenc, and H. Arslan. "A Survey on Multicarrier Communications: Prototype Filters, Lattice Structures, and Implementation Aspects," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 3, Mar. 2014, pp. 1312–38.

[15] K. Tsagkaris *et al.*, "Emerging Management Challenges for the 5G Era: Multi-Service Provision through Optimal End-to-End Resource Slicing in Virtualized Infrastructures Problem Statements and Solution Approaches," *Proc. IEEE/IFIP Network Operations Management Symp.*, Istanbul, Turkey, 2016, pp. 1297w–1300.

Biographies

Lei Zhang received his Ph.D. from the University of Sheffield, United Kingdom. He worked as a research engineer at Huawei and a research fellow in the 5G Innovation Centre (5GIC), Institute of Communications (ICS), University of Surrey, United Kingdom. He is now a lecturer at the University of Glasgow. His research interests broadly lie in communications and array signal processing, including physical layer network slicing (or RAN slicing), new air interface design (waveform, frame structure, etc.), the Internet of Things (IoT), multi-antenna signal processing, cloud radio access networks, massive MIMO systems, full-duplex, and so on. He holds 16 international patents on wireless communications.

Ayesha Ijaz received her B.Eng in electronic engineering from the University of Engineering & Technology, Taxila, Pakistan, in 2006, and her M.Sc and PhD. in mobile and satellite communications from the University of Surrey in 2008 and 2011, respectively. She is currently a research fellow at the Institute for Communication Systems (ICS), home of 5GIC at the University of Surrey. Her research interests include statistical signal processing and air interface design for next generation wireless communication systems.

Pei Xiao received his B.Eng, M.Sc., and Ph.D. degrees from Huazhong University of Science & Technology, Tampere University of Technology, and Chalmers University of Technology, respectively. Prior to joining the University of Surrey in 2011, he worked as a research fellow at Queen's University Belfast and held positions at Nokia Networks in Finland. He is a reader at the University of Surrey and also the technical manager of 5GIC, leading and coordinating research activities in all the work areas in 5GIC (http://www.surrey.ac.uk/5gic/research). His research interests and expertise span a wide range of areas in communications theory and signal processing for wireless communications.

Rahim Tafazolli is a professor and the director of ICS and 5GIC, University of Surrey. He has published more than 500 research papers in refereed journals and international conferences, and as an invited speaker. He is the editor of *Technologies for Wireless Future* (Wiley; Volume 1, 2004; Volume 2, 2006). He was appointed as a Fellow of the Wireless World Research Forum in April 2011, in recognition of his personal contribution to the wireless world. He also heads one of Europe's leading research groups.

ACCEPTED FROM OPEN CALL

Multi-Antenna Beamforming Techniques in Full-Duplex and Self-Energy Recycling Systems: Opportunities and Challenges

Duckdong Hwang, Sung Sik Nam, and Janghoon Yang

ABSTRACT

Full-duplex technique is considered as an essential component of the coming 5G wireless systems. Since the SI from its transmit antennas dominates over the signal received from remote sites, the suppression of SI is the first priority challenge in the design of FD systems, and thus combinations of analog SI cancellation with radio frequency domain approaches are considered. Applying BF techniques with multiple antennas on top of these combined approaches can further strengthen the SI suppression capability so that wireless systems can enjoy the benefit from FD operation without much concern on SI. However, various factors affect the design of BF in FD networks, and hence they provide a set of research challenges. Also, S-ER is an alternative way of utilizing SI in FD networks, where the RF radiation from a node is reused as an energy source. In this article, we go through these opportunities and challenges along with a survey of technical results developed so far in regard to the BF for handling the SI of FD systems. We look at various network models, the strategies for handling the SI, relay protocols, and antenna structure to see how these provide technical challenges for the BF design.

Full-duplex is considered as an essential component of the coming 5G wireless systems. Since the SI from its transmit antennas dominates over the signal received from remote sites, the suppression of SI is the first priority challenge in the design of FD systems, and thus combinations of analog SI cancellation with radio frequency domain approaches are considered.

INTRODUCTION

To meet the ever increasing demand for higher wireless data rate as well as support the massive number of devices required for fifth generation (5G) wireless systems, various techniques are being considered. In theory, full duplex (FD) systems double the throughput of half duplex (HD) wireless systems [1, 2] since they can accommodate the transmission and reception of a node in the same period to save spectral resource. Therefore, they can be very powerful tools for enhancing the wireless throughput of small cell networks or of device-to-device communication applications, where the reuse rate of spectral resource is increased many times [3, 4].

It has been shown that a large portion of the strong self-interference (SI) of FD systems can be suppressed by joint approaches of RF domain and baseband analog cancellation [5]. On top of these expensive hardware approaches, beamforming (BF) with multiple antennas at the FD nodes [6] can suppress residual interference so that the resulting FD systems are almost SI-free [7]. Multiple-input multiple-output (MIMO) BF has been utilized in various wireless network models to provide BF gain, spatial multiplexing, and interference suppression. Since FD can also be applied to various network models, SI suppression is an additional requirement to the original BF role in the model. Therefore, the BF design for FD should take the BF strategy, the protocols in relay networks, and various system parameters such as the number of antennas and spectral bandwidth into account, while SI suppression puts a constraint on the design of BFs.

Instead of suppressing the SI, we may turn it into an opportunity of harvesting energy from the SI [8, 9] to charge the batteries. Here, the nodes implement the data transmission and RF energy (their own) harvesting in the same period; otherwise, they need separate periods for these two operations. This type of FD mode is called self-energy recycling (S-ER), where the reception circuit of an FD node is replaced with an energy harvesting circuit with a necessary change of protocol. Since energy harvesting circuits are much simpler than those for SI suppression at FD nodes, this approach is cost efficient while we trade spectral efficiency (SE) of FD with energy efficiency (EE) of S-ER. Here, the requirements on MIMO BF for S-ER are opposite from those of the conventional approaches to improve SE since S-ER utilizes the SI instead of suppressing it. FD and S-ER can be jointly implemented at a node, and the RF energy from the receive antennas can be divided into two corresponding circuits.

In this article, we discuss opportunities and challenges in the applications of BF for the FD and S-ER networks with a survey of recent advances. Through the relay channel model, we show how the two approaches to SI (suppression and recycling) affect the BF design with additional constraints on the original design purposes. Beyond these example cases, various network models with different options (relay protocols, size of antenna number, wideband transmission, etc.) are also affected by the newly added constraints, and thus they harbor a rich set of research challenges.

Digital Object Identifier:
10.1109/MCOM.2017.1601029

Duckdong Hwang is with Konkuk University; Sung Sik Nam (corresponding author) is with Korea University; Janghoon Yang is with Seoul Media Institute of Technology.

For the adoption of FD technology in the coming 5G and beyond 5G wireless systems, these challenges should be resolved.

Basic Network Models

In Fig. 1, the FD MIMO system and three basic network models are depicted. The main interest of this article is the BF at the FD node, and hence multiple antennas are required at the FD node. Strong loopback channel (LBC) from the transmitter to the receiver at the same node brings the technical challenge of SI suppression. We denote this channel $\mathbf{H}_{RR}$ in the figure. Since the distance from these two antenna sets is much closer than that from a remote site, the SI through the LBC (typically the SI power is over 100 dB stronger) dominates the desired signal from remote sites. There have been joint efforts in RF domain and analog interference cancellation approaches, where near 100 dB interference suppression is achieved [1, 3] so that the residual SI signal after the suppression is shrunk to the dynamic range of the digital circuit for further processing. On top of these hardware approaches, the digital BF (the topic of this article) further reduces the SI sufficiently while serving the original role of the BF.

In Fig. 1b, three basic network models with FD capability are shown, although many derivatives of these basic models are possible. The power of FD comes from obviating separate physical channels for uplink and downlink at the expense of more hardware and digital BF for SI suppression. In the multiuser BS channel model of Fig. 1b-1, the base station (BS) adopts FD to accommodate the uplink and downlink in a single period so that the BS can decode the uplink transmission while it sends the downlink signals toward its subordinate terminals. In the bidirectional FD of the device-to-device (D2D) communications model in Fig. 1b-2, a pair of mobile devices borrow spectral resources from a BS and exchange message signals. Finally, the FD relay in the model of Fig. 1b-3 relays the BS downlink signal to the serving mobile terminal (the reverse direction in the uplink). Multiple FD relays can be installed in a cellular network, as in Fig. 1c, where there are three BSs and three relays. Among a set of relay protocols, this article sees the impact of FD on the two most popular protocols, decode-and-forward (DaF) and amplify-and-forward (AaF).

In practice, we assume that various BSs (macro, pico, femto, and fixed relay) are attached to power grids and that they do not suffer from power shortage. On the other hand, mobile terminals and relays rely on battery power with limited capacity, so the EE of these devices is of practical importance. S-ER [8, 9] is a strategy to improve the EE of these devices by restoring a portion of its transmit energy. As small cell networks dominate in 5G systems and the distances between network nodes shrink, the chance of harvesting and recycling the RF energy becomes more and more practical. Although S-ER can be used for mobile terminals, Fig. 2 shows the RF-powered S-ER relay network and associated data and energy transfer protocol along with some example energy signal spectra, where signals with high peak-to-average-power ratio such as white noise and chaotic signals are recommended for energy conversion efficiency. As shown in Fig. 2b, the slot structure of an S-ER relay network saves one slot period compared to the case without S-ER. Here, RF powering from a BS is included since the recycled energy is just a portion of transmit energy and thus is not enough to extend the battery lifetime sufficiently. The data is delivered from the BS to an Rx node via the relay during the two time slots while the energy is harvested in the second slot. The relay signal from the LBC ($\mathbf{H}_{RR}$) and the energy signal from the BS add up at the receive antenna of the energy harvest circuit. Depending on the protocol, the ratio of two time slot lengths τ/T can be adjusted. In contrast to the FD relay, where data relaying and data reception share the same time period, data relaying and energy transfer share the second slot in the S-ER relaying. Otherwise, we need an additional time slot to separate the data relaying and energy harvesting. It is the slot structure of FD and S-ER systems

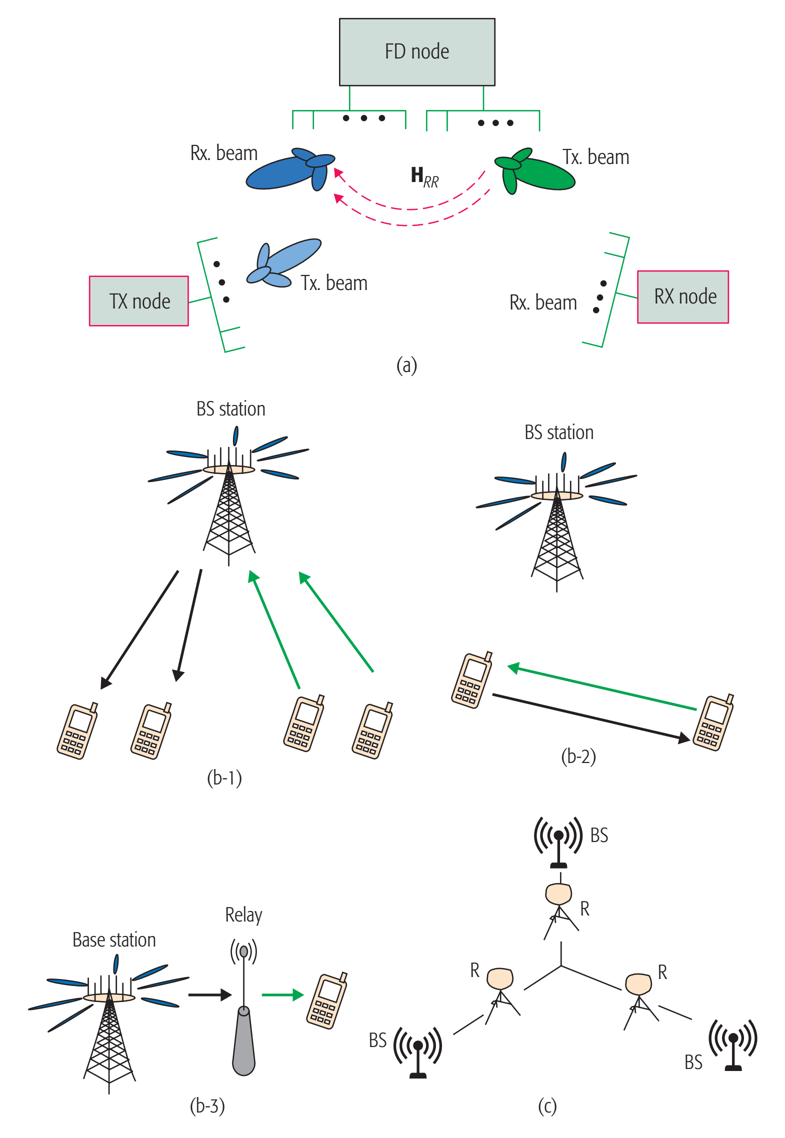


Figure 1. a) Full-duplex MIMO system with multi-antenna transmission and reception nodes; b) three basic network models where FD techniques can be applied; b-1) FD base station in the multiuser channel model; b-2) bidirectional FD in the device-to-device (D2D) communications model; b-3) FD relay channel; c) a simple cellular model with three base stations and three relays.

The FD and S-ER can be jointly applied when two circuits of FD and S-ER are connected to the receive antennas of a relay node. In this case, the energy from the RF circuit is divided into two circuits so that the relay can forward the message signal and harvests a portion of RF energy into the battery at the same time.

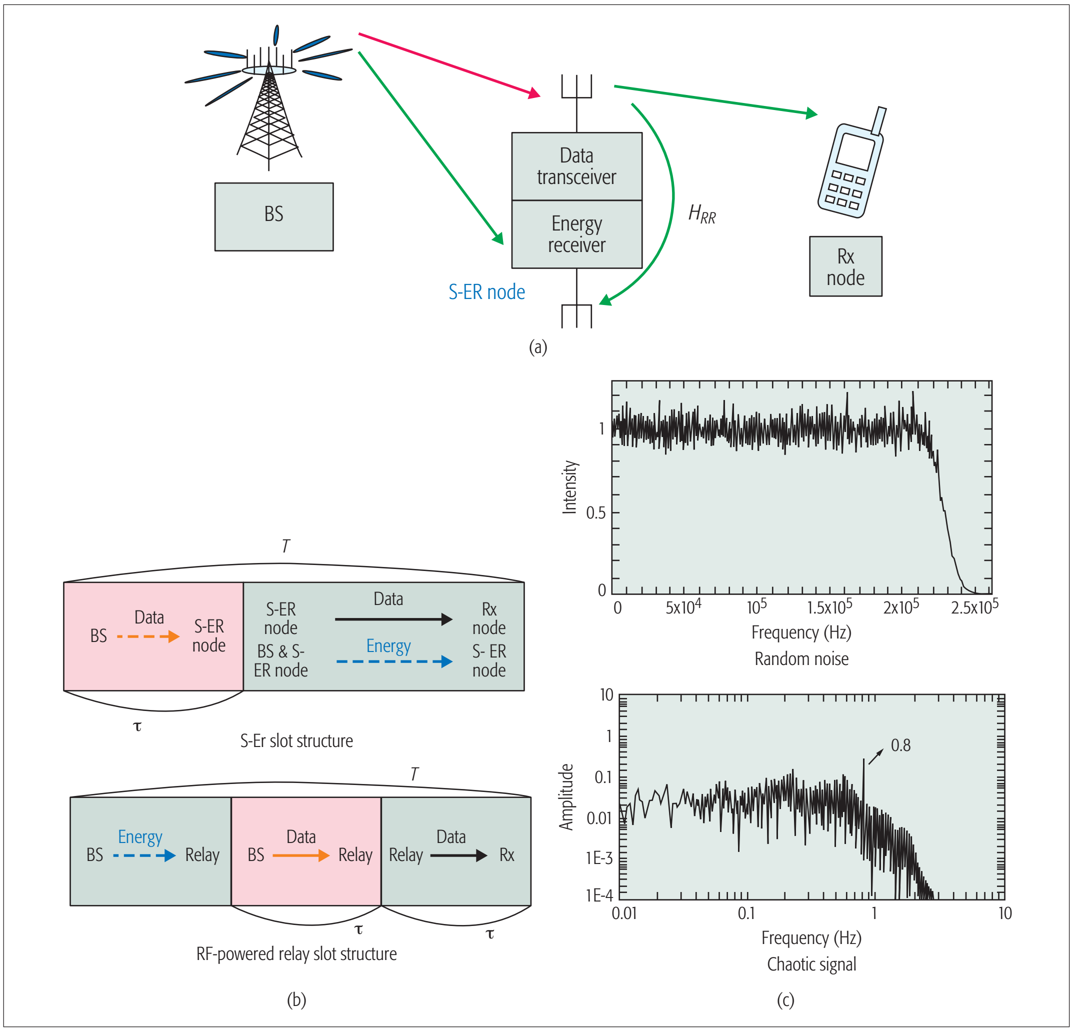


Figure 2. a) The RF powered self-energy recycling (S-ER) relay network; b) top: the data and energy transfer protocol of S-ER, where the red period corresponds to the first slot transmission and the green period corresponds to the second slot transmission; bottom: the data and energy transfer protocol of an RF-powered relay network; c) the spectra of energy signals.

that improves the SE. On the other hand, the S-ER system can improve the EE by recycling its own transmit energy. The FD and S-ER can be jointly applied when two circuits of FD and S-ER are connected to the receive antennas of a relay node. In this case, the energy from the RF circuit is divided into two circuits so that the relay can forward the message signal and harvest a portion of RF energy into the battery at the same time. Since the SI impacts on the two receiver circuits are so contrasting, these circuits should be designed accordingly so that the SI should be sufficiently suppressed (in FD) or survive the path loss (PL) (in S-ER). As an example, the S-ER receive antenna can be placed closer to the transmit antennas while the FD receive antennas are separated as much as possible.

The structure of FD nodes differs in baseband (BB) operations while their transmit and receive circuits share common structures. For transmit and receive sides, there are separate RF chains and baseband beamformers. DaF and AaF operations are used for the FD relay networks. A DaF relay terminates the first hop transmission and sends the re-encoded message signal to the beamformer only if it succeeds in decoding the first hop signal, while an AaF relay simply scales the received signal according to the transmit power requirement and sends it to the beamformer. In FD BSs and bidirectional D2D devices, the message reception is terminated with decoding, and a new set of messages are encoded to be sent to the transmit BF.

Interference Suppression BF

It is the existence of LBC ($\mathbf{H}_{RR}$) at an FD node that brings the challenge of SI suppression and the opportunity of S-ER. The performance and properties of FD BFs vary according to the statistical characteristics of $\mathbf{H}_{RR}$ as well as the strategy (to suppress or to harvest energy) of the FD node. According to the node strategy, the antenna set pair (transmit set and receive set) design allows the line-of-sight (LOS) component or blocks it. Also, the performance of the analog domain cancellation circuits affects the strength of the LOS component. When the LOS component dominates, the rank of channel matrix is almost one, while the local scatterers around the antenna set pair of an FD node [5] contribute to the non-LOS components of $\mathbf{H}_{RR}$ and make the rank of channel matrix close to full.

Zero Forcing

Zero forcing beamforming (ZFBF) is an intuitive criterion to suppress SI at an FD node. Let the transmit BF matrix be $\mathbf{W}_t$ and the receive BF matrix be $\mathbf{W}_r$; then the SI can be written as $\mathbf{W}_r\mathbf{H}_{RR}\mathbf{W}_t$. The condition where this matrix product of effective SI channel is forced into the all zero matrix is called zero forcing condition (ZFC) [7, 10]. All pairs (W_r, $\mathbf{W}_t$) satisfying the ZFC constitute the ZFBF solution set, where the optimization of ZFBF within the set is possible [10] for FD AaF relay systems with single-stream transmission. Once $\mathbf{W}_r$ (or $\mathbf{W}_t$) is fixed (typically toward the directions of singular vectors of the channel matrix that the receive, or transmit, antennas see), the BF $\mathbf{W}_t$ (or $\mathbf{W}_r$) can be found from the projection onto the orthogonal space of $\mathbf{W}_r\mathbf{H}_{RR}$ (or $\mathbf{H}_{RR}\mathbf{W}_t$)[7, 10].

Since ZFBF reveals the insight on the spatial degree of freedom (DoF) spent for suppressing the SI, the available DoF after SI suppression can be determined. Let N_t and N_r be the numbers of antennas at the transmit side and receive side of the FD node, respectively, and N ($N \le \min(N_t, N_r)$ be the number of data streams. When we fix $\mathbf{W}_t$, the dimension of the orthogonal space of $\mathbf{H}_{RR}\mathbf{W}_t$ is $N_r - N$ with probability one, given that $\mathbf{H}_{RR}$ is an independent and identically distributed (i.i.d.) matrix. The N rows of $\mathbf{W}_r$ should lie in the orthogonal space of $\mathbf{H}_{RR}\mathbf{W}_t$, and thus $N \le N_r - N$ ($N \le N_r/2$). Similarly, from fixing $\mathbf{W}_r$, we can find $N \le N_t/2$. We can select to fix $\mathbf{W}_r$ or $\mathbf{W}_t$ such that

$$N \le \frac{\max(N_t, N_r)}{2}$$

streams (or DoF) are supported at a specific FD node. Note that the full rank assumption of $\mathbf{H}_{RR}$ is quite strong and corresponds to one of the worst case scenarios. When $\mathbf{H}_{RR}$ is rank-deficient, we can utilize it into the DoF gain as well. Note also that ZFBF keeps the same form regardless of BB operations.

MMSE and Throughput Maximization

For the minimum mean square error (MMSE) criterion and the capacity (throughput) maximization, finding the correlation matrix of interference plus noise at the receive antennas of an FD node is the key step. Except for the FD relay with AaF protocol, a link terminates after the receive BF of an FD node and another link starts before the transmit BF of the FD node. Therefore, the design of $\mathbf{W}_r$ and $\mathbf{W}_t$ can be separated to optimize each link. However, the convenience of design separation is not possible in the AaF relay network. The received signal vector of the FD relay with AaF protocol is given as

$$\mathbf{y}_r(t) + \sum_{i=0}^{\infty}(\mathbf{H}_{RR}\mathbf{W})^i[\mathbf{H}_1\mathbf{X}(t - i\tau) + \mathbf{n}(t - i\tau)]. \quad (1)$$

Here, $\mathbf{W} = \mathbf{W}_t\mathbf{W}_r^H$, τ is the processing delay at the AaF FD relay, $\mathbf{H}_1$ is the first hop channel matrix, $\mathbf{X}(t)$ is the BS transmit signal at time t, and $\mathbf{n}(t)$ is the additive white Gaussian noise with variance σ^2.

The signal term is $\mathbf{H}_1\mathbf{X}(t)$, and all other terms are SI terms with the additive noise terms. Then the correlation matrix of the SI plus the noise becomes a matrix power series as $\mathbf{R}_y = \Sigma_{i=1}^{\infty}(\mathbf{H}_{RR}\mathbf{W})^i[H_1\mathbf{R}_X\mathbf{H}_1^H + \sigma^2\mathbf{I}](\mathbf{W}^H\mathbf{H}_{RR}^{\boldsymbol{H}})^i + \sigma^2\mathbf{I}$ when $\mathbf{R}_X$ is the correlation matrix of $\mathbf{X}(t)$. Given that $\mathbf{W}$ is power limited (the Frobenius norm of $\mathbf{W}$ is limited) and the singular values of $\mathbf{H}_{RR}$ are sufficiently small ($\ll 1$) due to PL, we can show that the correlation matrix converges. Consider the matrix property $(\mathbf{I} - \Sigma)\mathbf{X}(\mathbf{I} + \Sigma)^H + (\mathbf{I} + \Sigma)\mathbf{X}(\mathbf{I} - \Sigma)^H = 2\Sigma\mathbf{A}\Sigma^H$, where Σ is an arbitrary diagonal matrix and $\mathbf{X} = \Sigma_{i=1}^{\infty}\Sigma^i\mathbf{A}(\Sigma^H)^i$. Then, using the eigen decomposition $\mathbf{H}_{RR}\mathbf{W} = \mathbf{S}\Lambda\mathbf{S}^{-1}$ with the eigen values $\lambda_1, \dots, \lambda_{N_r}$, we can show that the correlation matrix converges to $\mathbf{S}\psi \odot \mathbf{S}^{-1}[\mathbf{H}_1\mathbf{R}_X\mathbf{H}_1^H + \sigma^2\mathbf{I}](\mathbf{S}^{-1})^H \odot \psi^H\mathbf{S}^H + \sigma^2\mathbf{I}$, where ψ is the matrix whose ith row and jth column element is

$$\sqrt{\frac{\lambda_i\lambda_j^*}{1-\lambda_i\lambda_j^*}}$$

and $\odot$ is the Hadamard product operator. The correlation matrix of the SI plus the noise at the mobile terminal can be found as $\mathbf{H}_2\mathbf{W}\mathbf{R}_y\mathbf{W}^H\mathbf{H}_2^H + \sigma^0\mathbf{I}$, where $\mathbf{H}_2$ is the second hop channel matrix. With the correlation matrix, we can optimize $\mathbf{W}$,$\mathbf{R}_X$ and the BF at the mobile terminal to minimize the mean square error (MSE) [11] or to maximize the total throughput [12]. Since the Hadamard product operations of the correlation matrix are hard to handle, we may resort to bounding or approximation techniques of the Hadamard operation. Also, rank one transmission cases simplify the operations involved with the Hadamard products.

Since we would have $\mathbf{R}_y = \mathbf{H}_1\mathbf{R}_X\mathbf{H}_1^H + \sigma^2\mathbf{I}$ without the LBC, the LBC transforms the correlation matrix with $\mathbf{S}\psi \odot \mathbf{S}^{-1}$ and adds $\sigma^2\mathbf{I}$. In AaF relays with HD, it has been shown that the beamformers at the relay are canonically decomposed such that its receive BF ($\mathbf{W}_r$) and transmit BF ($\mathbf{W}_t$) are composed of the left singular vectors of the first hop and the right singular vectors of the second hop, respectively. Then the resulting end-to-end (E2E) channels are parallelized so that water-filling-like power allocations optimize the throughput [11, 12]. Certainly, this neat and easy decomposition should not be the case in AaF FD relays since the above transformation makes the BF optimization more challenging and involved. There can be other BF design metrics such as signal-to-interference-plus-noise power ratio (SINR) and some derivatives of SINR in the multiuser context. With the AaF protocol, SINR also suffers from infinite feedback of SI terms, and we may take a similar approach to the matrix power series convergence.

Protocol Impact

The relay protocol also affects the design of the FD relay beamformers. While the SI is fed back repeatedly in the AaF protocol, it affects the DaF relay decision on the message signals once, and its effect no longer propagates into future signals. Hence, the SI contributes to the outage (or the symbol error probability) at the relay, and this outage affects the E2E performance. To see what the impact of relay protocols is on the throughput of the FD relay networks, we conduct some simulations, the results of which are shown in Fig. 3. In these simulations, a single stream transmission is assumed for simplicity (the rank one transmission case of $N = 1$). We compare the average throughput of AaF FD relay and DaF FD relay with the two ZFBF schemes (depending on whether the

The relay protocol also affects the design of the FD relay beamformers. While the SI is fed back repeatedly in the AaF protocol, it affects the DaF relay decision on the message signals once and its effect no longer propagates into the future signals. Hence, the SI contributes to the outage at the relay and this outage affects the E2E performance.

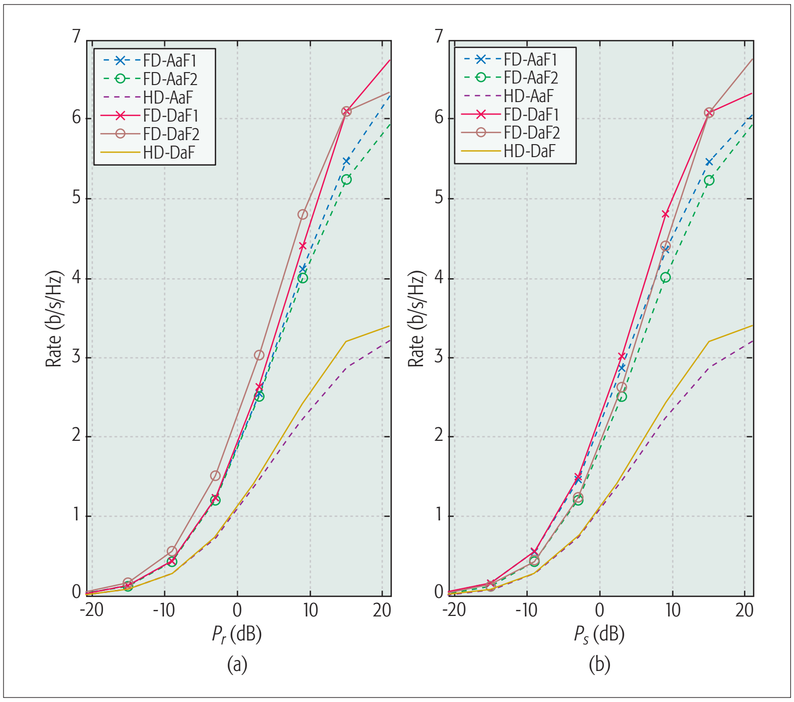


Figure 3. The average rate comparison of AaF FD relay system and DaF FD relay system with two ZFBF schemes when the second hop SNR P_r varies and the first hop SNR P_s is fixed to 9 dB. Here, $N_t = N_r = 4$.

transmit or receive BF is directing in a single channel direction) when the source power P_s is fixed (Fig. 3a) and the relay power P_r is fixed (Fig. 3b). Note that the throughput rates of the two relay protocols with HD are plotted as well for reference.

When we increase the second hop SNR P_r, the performance gaps of the two protocols narrow down to near zero as in Fig. 3a. Thus, the use of a simpler AaF protocol FD relay can be advocated for strong second hop link quality. In HD relay networks, the DaF protocol gives better performance than the AaF protocol when the first hop link quality is strong. In the simulations of FD ZFBF schemes, the DaF protocol outperforms the AaF protocol consistently, although the gain gets wider as the first hop SNR P_s increases in Fig. 3b. Thus, the complex hardware for DaF relay is worth the price in the general FD relay scenarios in stark contrast to the HD relay channel. The first ZFBF scheme aligns the direction of $\mathbf{W}_r$ with the direction of the first hop channel vector h1 and finds $\mathbf{W}_t$ from the direction of the second hop channel vector h2 projected to the orthogonal space to the rows of $\mathbf{W}_r\mathbf{H}_{RR}$. Thus, the zero forcing operation is done in the second hop channel in the first scheme, while the second ZFBF scheme does the operation in the first hop channel. The first ZFBF scheme suffers more when the second hop link quality is weak (in Fig. 4a) contrary to the second ZFBF scheme (in Fig. 4b).

To see the impact of MIMO FD relay in the cellular networks, a simple system-level simulation was executed for the cellular network model of Fig. 1c. It is assumed that three BSs being equidistant from each other with the distance of 1 km and 30 randomly distributed mobile stations (MSs) are placed within a 1 km^2 square. For simplicity, a single antenna at both MSs and BSs, an urban-micro PL model following the Third Generation Partnership Project (3GPP) spatial channel model (SCM), and round-robin scheduling are assumed. In an HD-relay system, signals from the BS at the first hop and the relay station (RS) at the second hop are combined with maximum ratio combining (MRC), while the MRC and maximum ratio transmission (MRT) are used for MIMO processing at HD relays, In an FD-relay system, signals from the BS and RS are naively added at the RF front-end of an MS while MRC and ZF transmit BF with respect to SI are employed. The average system spectral efficiencies of the FD-MIMO relay system are compared to those of the HD-MIMO relay system in Fig. 4. Three RSs were placed symmetrically with one for each cell with distance relative to cell radius. The FD relay system provides larger average SE by 30–65 percent. However, it is noted that gain in the SE tends to decrease as the relays are positioned further away from the associated BS since as the FD relay moves toward the cell boundary, it relays more interference from other BSs and other RSs. The result suggests that more advanced schemes need future research to improve the system performance with the cell edge FD relays.

Self Energy Recycling BF

In the S-ER strategy of FD structures, they extend battery lifetime by harvesting the energy of SI. SI is welcome in the S-ER strategy rather than an annoying problem to be suppressed. Thus, it is better to minimize the PL between two antenna sets through enforcing LOS or reducing the distance between them. Also, redirecting the transmitter leakage power in the circuit to the battery can boost the S-ER performance. The signal an S-ER relay receives in the first protocol slot is the same as that an HD relay receives in the first slot. Not only does an S-ER relay relay the received signal toward the Rx node in the second slot, but its energy receiver also harvests the RF energy from the feedback signal and the signal from the BS as in Fig. 2. Two input signals to the energy receiver can be added constructively to improve the S-ER performance [8]. The signal the S-ER relay receives in the second slot is given as

$$\mathbf{y}_r(t) = \mathbf{H}_1\mathbf{W}_B\mathbf{x}_B + \mathbf{H}_{RR}\mathbf{W}[\mathbf{y}_r(t-\tau) + \mathbf{n}(t-\tau)] + \mathbf{n}(t).$$

Here, $\mathbf{W}_B$ and $\mathbf{x}_B$ are the BS energy BF and BS energy signal, respectively.

The beamformers $\mathbf{W}_B$, $\mathbf{W}$ and the BS energy signal power $E[\|\mathbf{x}_B\|^2]$ should be optimized to maximize the E2E performance while the energy requirement at the S-ER node is met. Among the many strategies on constraining the energy requirement at the S-ER relay, the strictest one is to keep the transmit power of the S-ER relay below the RF harvested energy to maximize the battery lifetime. The S-ER relay transmit power is $\mathrm{E}_n\,[Tr[\mathbf{W}\mathbf{y}_r\mathbf{y}_r^H\mathbf{W}^H]]$ and the harvested energy is $\xi\mathrm{E}_{n_R}\,[\|\mathbf{H}_{RR}\mathbf{W}\mathbf{y}_r\}\|^2] + \xi\|\mathbf{H}_1\mathbf{W}_B\|_2^2\|\mathbf{x}_B\|^2$ with the matrix Frobenius norm $\|\cdot\|_2^2\|$ and the energy conversion efficiency $\xi(\leq 1)$. Setting the relay transmit power to be less than the harvested power makes a constraint element, while various E2E performance metrics can be the objective of optimization problems. For example, we may maximize the E2E signal-to-noise ratio (SNR) when single stream transmission ($N = 1$) is considered [9].

The authors in [9] show that the optimal rank one $\mathbf{W}_B$ should be steered in the direction of the strongest right singular vector of $\mathbf{H}_1$, and the optimal $\mathbf{W}$ should be an outer product of rank one vectors as $\mathbf{W} = \mathbf{w}_t\mathbf{w}_r^H$, where $\mathbf{w}_r$ is steered in the direction of the strongest left singular vector of $\mathbf{H}_1$ and w_t is a vector residing in the space composed of the strongest right singular vectors of $\mathbf{H}_2$ and $\mathbf{H}_{RR}$. Exact direction and the magnitude of $\mathbf{w}_t$ can be found by a geometric geodesic approach, where an angle-search-based optimization gives simple closed form expressions of them. Given the set of optimal BFs, there remain two more factors affecting the E2E SNR. One is the BS power allocation between the first slot power (P_1 for data transmission) and the second slot power (P_2 for energy transfer), where the case with ($P_2 \gg P_1$) gives superior E2E SNR to the opposite case. Also, for very large values of P_2, the E2E SNR of S-ER relaying and that of conventional relaying with the same relay power converge. The other factor affecting the E2E SNR is the energy recycling ratio defined as the product of energy conversion efficiency (ξ) and the PL between the transmit and receive antenna sets. As the energy recycling ratio approaches one, the BF $\mathbf{w}_t$ tries to make a balance between two directions, one toward the channel direction to the Rx node and the other toward the strongest right singular vector direction of $\mathbf{H}_{RR}$ so that the Rx SNR can be maximized with the energy constraint being satisfied. When the energy recycling ratio approaches zero, $\mathbf{w}_t$ points in the Rx channel direction to maximize the E2E SNR while relying mostly on the RF energy from the BS.

In Fig. 5, the rates of an S-ER-based AaF relay network in Fig. 2a with different hop distance pairs (first hop distance $D1$, second hop distance $D2$) against the S-ER ratio appear. At high values of S-ER ratio, the rates are noticeably boosted while the effect of S-ER is merely at low S-ER ratio. All the pairs of rate curves show that the power difference of two time slots (P_1 and P_2) is an important factor in the rate performance.

Massive MIMO, Small Cells, and Wideband

Extending the number of antennas of MIMO systems to very large numbers provides many benefits. The strong directivity of massive MIMO systems helps FD massive MIMO systems to suppress the SI while the strong BF gain from the massive antennas can be helpful for S-ER. With an ideal assumption of the law of large numbers (LLN), the inner product of two independent channel vectors converges to zero, which can simplify the BF design for the SI suppression in FD systems since it is sufficient to form the BF direction toward the Rx node. On the other hand, the exploitation of SI requires more consideration in the BF design since it is challenging to meet the two objectives of the maximization of E2E SNR and the S-ER requirement with a single BF. While the LLN holds with the infinite number of antennas, there is still the issue of BF design even for the FD nodes in practical massive MIMO systems, where the numbers of antenna elements are still finite.

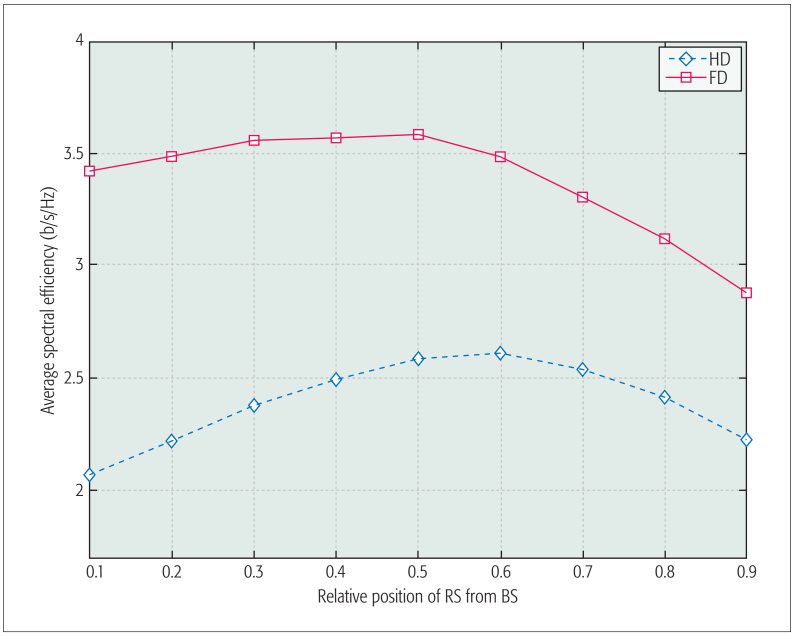


Figure 4. The average system spectral efficiencies of the HD-MIMO relay and the FD-MIMO relay of three cell networks. Here, the relays have two transmit antennas and two receive antennas.

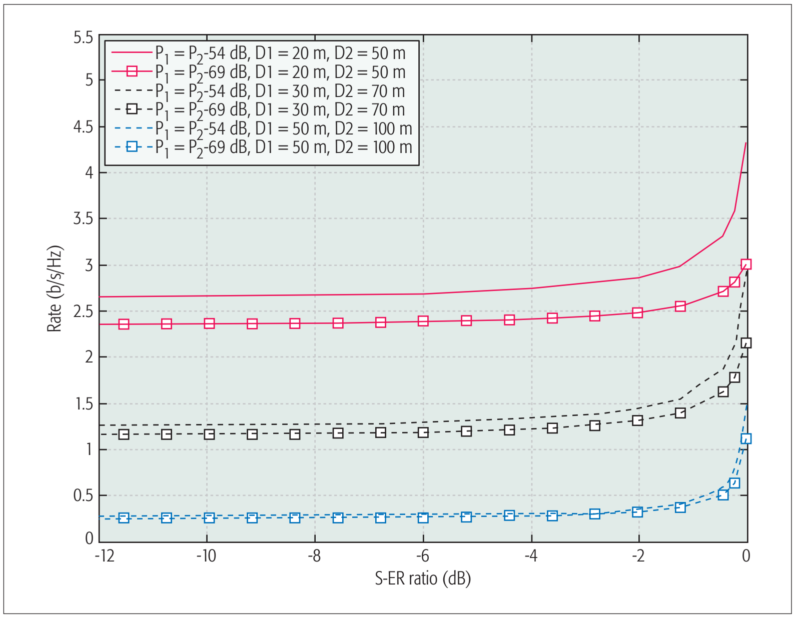


Figure 5. The rates of S-ER based AaF relay network against the S-ER ratio with different hop distance pairs. Here, the time slot structure is given as $\tau/T = 0.5$ and the numbers of antennas at the S-ER relay are given as $N_t = N_r = 4$. The BS power of 23 dBm is uniformly spread over a 10 MHz bandwidth and the PL exponent is set to two.

Another major direction of 5G system evolution is small cells with heterogeneous network architectures. Dense cells populate the service area with excessive cross- and inter-tier interferences. Adding FD nodes in such networks may aggravate the interference situation due to the coexistence of uplink and downlink in the same channel. The beamformers at an FD node should fight not only for the SI but also for the cross- and inter-tier interferences. Unlike FD nodes in small cell networks, the S-ER in small cells neither impairs the interference situation much nor

Interference handling approach	SI suppression, self-energy recycling
Network channel model	Multiuser channel, bidirectional communication, one-way relay, two-way relay.
Performance metric	Zero forcing, MMSE, capacity, SNR, SINR.
Relay protocol	Decode-and-forward, amplify-and-forward.
Technical trends	Massive MIMO, small cells, wideband, bandwidth aggregation.
Performance analysis	Throughput, outage.

Table 1. Factors affecting the BF design for FD systems.

invokes additional challenges since it only harvests the RF energy. Instead, the S-ER can take the cross- and inter-tier interference into energy sources in addition to the SI.

Modern wireless access systems operate on wide bandwidths, and orthogonal frequency-division multiplexing systems are prevalent. These wideband systems provide additional frequency space for the optimization of FD systems. Some frequency bands may provide better channel sets for the FD systems and others may have better channel sets for S-ER systems. Joint optimization of frequency band and BF allows further enhancement of these systems. The idea can be extended to the resource allocation approach where some frequency bands are reserved for the FD operation or for the S-ER operation. These approaches may be well suited for systems with joint implementation of FD and S-ER. Some wireless standards such as LTE support bandwidth aggregation where multiple carriers at different frequencies operate as a single virtual channel. When carriers are scattered widely over frequency bandwidth, the FD system is likely to have a similar benefit to that of wideband systems.

Research Challenges

There are lots of opportunities and challenges in the BF design for FD systems depending on various factors as summarized in Table 1. The most important factor is the approach to SI, where the SI is to be suppressed (FD) or recycled as an energy source (S-ER). If the FD operation is chosen, the SI should be suppressed at the FD node (ZFBF), or the effect of SI on the final E2E performance metric should be taken into account in designing the FD BF. However, in S-ER, the SI itself never affects the E2E performance, and thus the BF can utilize it as an energy source as long as the E2E performance is satisfied. Various network models have been developed, and we may apply FD BF and/or S-ER BF in those network models to improve performance. More variations of relay networks than those shown in Table 1 are possible if we take the multiuser cases into account.

Various performance metrics can be considered. The ZFBF approach is intuitive and gives insight such as DoF of the network. We may work with MMSE or the throughput (capacity) to optimize the BF. The SNR or SINR can be used in multiuser networks. The relay protocol affects much of the BF design since the first hop transmission is terminated at the DaF relay, and the two hop transmission can be separated. New technical trends in 5G and beyond 5G wireless systems such as massive MIMO, small cells, and wideband transmission affect the BF design. Also, performance analysis, such as outage probability and throughput for each BF design, is an interesting research direction.

Conclusion

Utilizing the SI cleverly in FD-like systems is an active research field for 5G and beyond 5G wireless systems, where the network elements become dense and the removal of the up- and downlink barrier (via FD) for SE becomes prevalent. We go through the impacts of BF with multiple antennas in various network models, where the cooperating BFs are employed at the network nodes to improve the SE and EE of FD and S-ER systems. In addition to the RF design and the analog circuits, digital BF strengthens the performance of these network models. It turns out that this area is a rich field because utilizing the SI renders abundant technical challenges for BF design. Several performance metrics can be considered for BF optimization, and the relay protocol affects BF design in relay networks.

Acknowledgment

This work is supported by the KU Research Professor Program of Konkuk University.

References

[1] X. Zhang *et al.*, "Full-Duplex Transmission in PHY and MAC Layers for 5G Mobile Wireless Networks," *IEEE Wireless Commun.*, 2015, vol. 22, no. 5, pp. 112–21.

[2] G. Liu *et al.*, "In-Band Full Duplex Relaying: A Survey, Research Issues and Challenges," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 2, 2015, pp. 500–24.

[3] S. Goyal *et al.*, "Full Duplex Cellular Systems: Will Doubling Interference Prevent Doubling Capacity," *IEEE Commun. Mag.*, vol. 53, no. 5, May 2015, pp. 121–27.

[4] L. Wang *et al.*, "Exploiting Full Duplex for Device-to-Device Communications in Heterogeneous Networks," *IEEE Commun. Mag.*, vol. 53, no. 5, May 2015, pp. 146–52.

[5] M. Heino *et al.*, "Recent Advances in Antenna Design and Interference Cancellation Algorithms for In-Band Full Duplex Relays," *IEEE Commun. Mag.*, vol. 53, no. 5, May 2015, pp. 91–101.

[6] T. Riihonen *et al.*, "Mitigation of Loopback Self-Interference in Full-Duplex MIMO Relays," *IEEE Trans. Signal Processing*, vol. 59, no. 12, 2011, pp. 5983–93.

[7] H. A. Suraweera *et al.*, "Low-Complexity End-to-End Performance Optimization in MIMO Full-Duplex Relay Systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, 2014, pp. 913–27.

[8] Y. Zeng and R. Zhang, "Full-Duplex Wireless-Powered Relay with Self-Energy Recycling," *IEEE Wireless Commun. Letters*, 2015, vol. 4, no. 2, pp. 201–04.

[9] D. Hwang *et al.*, "Self-Energy Recycling for RF-Powered Multi-Antenna Relay Channels," *IEEE Trans. Wireless Commun.*, vol. 12, no. 2, 2017, pp. 812–24.

[10] D. Hwang et al., "Optimization of Zero Forcing Beamfomer for the Full Duplex Relay System," *IEEE Commun. Letters*, vol. 20, no. 8, 2016, pp. 1583–86.

[11] W. Guan and H. Luo, "Joint MMSE Transceiver Design in Non-Regenerative MIMO Relay Systems," *IEEE Commun. Letters*, 2008, vol. 12, no. 7, pp. 517–19.

[12] X. Tang and Y. Hua, "Optimal Design of Non-Regenerative MIMO Wireless Relays," *IEEE Trans. Wireless Commun.*, vol. 6, no. 4, 2007, pp. 1398–1407.

Additional Reading

[1] S. Huberman and T. Le-Ngog, "Full-Duplex MIMO Precoding for Sum-Rate Maximization with Sequential Convex Programming," *IEEE Vehic. Tech.*, vol. 64, no. 11, 2015, pp. 5103–12.

[2] D. Nguyen *et al.*, "Precoding for Full Duplex Multiuser MIMO Systems: Spectral and Energy Efficiency Maximization," *IEEE Trans. Signal Processing*, 2013, vol. 61, no. 16, pp. 4038–50.

[3] U. Ugurlu, T. Riihonen, and R. Wichman, “Optimized In-Band Full-Duplex MIMO Relay under Single-Stream Transmission,” *IEEE Vehic. Tech.*, 2016, vol. 65, no. 1, pp. 155–68.

Biographies

Duckdong Hwang [M’05] received his B.S. and M.S. in electronics engineering from Yonsei University, Korea. He received his Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, in May 2005. In 2005, he joined Digital Research Center, Samsung Advanced Institute of Technology as a research staff member. Since 2012, he has been a research associate professor in the School of Information and Communication Engineering at Sungkyunkwan University and in the Electrical Engineering Department of Konkuk University, Korea. He also worked for Daewoo Electronics in Korea from 1993 to 1998 as an engineer. He is interested in the physical layer aspect of the next generation wireless communication systems.

Sung Sik Nam [S’05, M’09] received B.S. and M.S. degrees in electronic engineering from Hanyang University, Korea, in 1998 and 2000, respectively. He received an M.S. degree in electrical engineering from University of Southern California in 2003, and a Ph.D. degree at Texas A&M University in 2009. From 1998 to 1999, he worked as a researcher at ETRI, Korea. From 2003 through 2004, he worked as a manager at the Korea Telecom Corporation, Korea. From 2009 to 2016, he was with Hanyang University and Sungkyunkwan University, Korea, respectively. Since 2017, he has been with Korea University. His research interests include the design and performance analysis of wireless communication system including wireless optical communication.

Janghoon Yang received his Ph.D. in electrical engineering from University of Southern California in 2001. He is currently an associate professor in the Department of New Media, Seoul Media Institute of Technology, Korea. From 2001 to 2006, he was with the Communication R&D Center, Samsung Electronics. From 2006 to 2010, he was a research assistant professor in the Department of Electrical and Electronic Engineering, Yonsei University. He has been with Seoul Media Institute of Technology since 2010. He has published numerous papers in the area of multi-antenna transmission, signal processing, and control. His research interests include wireless systems and networks, artificial intelligence, control theory, neuroscience, affective computing, and intervention for special education.

Toward Context-Aware Mobile Social Networks

Zhiyong Yu, Daqing Zhang, Zhu Wang, Bin Guo, Ioanna Roussaki, Kevin Doolin, and Ethel Claffey

The authors classify CA-MSNs into four categories, and divide their life cycle into four phases: discovery, connection, interaction, and organization. They then introduce personal and community context, and discuss the corresponding taxonomy. They also discuss how such context can be leveraged to enhance each life cycle phase.

Abstract

CA-MSNs are more intelligent and user-friendly than conventional online or mobile social networks. We first classify CA-MSNs into four categories, and divide their life cycle into four phases: discovery, connection, interaction, and organization. We then introduce personal and community context, and discuss the corresponding taxonomy. Subsequently, we elaborate how such context can be leveraged to enhance each life cycle phase. We also present our practices on designing various CA-MSN applications. Finally, future research directions are identified to shed light on the next generation MSNs from the context awareness perspective.

Introduction

Mobile social networks (MSNs) are becoming killer applications that can show the power of combining mobile computing with social networking [1]. MSNs are not only an elementary extension of existing online social networks (i.e., conventional MSNs), but also *revolutionizing social networks by bringing anywhere anytime social interaction with higher-level intelligence*. The former is reached by smart mobile phones' inherent property via wireless communication, whereas the latter is enabled through utilizing the comprehensive users' context acquired or inferred from fertile data sources such as web services, social networking sites, wearable/mobile devices, and environmental wired/wireless sensor networks [2, 3]. We name the latter context-aware mobile social networks (CA-MSNs).

Although online social networking services have been very successful in attracting billions of users to socialize in cyber space in a short time, none of them tap into the vast amount of context affiliated with users who shuttle constantly across the physical and virtual worlds with feature-rich smartphones. Nowadays there is an unparalleled chance to comprehensively understand the context surrounding individuals or communities in almost any scene [2]. Motivated by this observation, we aim to exploit new facets of context that are vital to MSNs, and examine how context awareness will shape the future MSN paradigm. More specifically, this article:

- Characterizes CA-MSNs by comparing them to conventional MSNs and provides a taxonomy for CA-MSNs with corresponding application scenarios
- Proposes a methodology of creating CA-MSN applications by investigating the usage of personal and community context in their lifecycle phases
- Designs three CA-MSN applications with guidance from the methodology, which shows the power of context awareness through our practices and outputs some visions of future MSNs

Characteristics and Taxonomy of CA-MSNs

An increasing number of people are socializing and grouping in cyber space by using online social networks regularly. With the quick penetration of sensor-equipped smartphones, social networking services (e.g., Twitter and Facebook) tend to create mobile phone applications, which can provide online users with "here and now" access from their smartphones. In turn, native MSNs (e.g., Foursquare) have been developed to construct communities for real-world mobile users. The line between social network services on the web and mobile applications is being blurred. As a result, two trends are joined: online social networks are extended for mobile access and localization through mobile phone browsers or applications, and native MSNs utilize user profiles, activities, and contents generated via online social networks. This way, MSNs can be maintained remotely and virtually just like traditional online social networks, and can also be leveraged to support face-to-face and spontaneous interaction. However, these two trends have different genes. Compared to conventional MSNs, CA-MSNs use rich and high-semantic-level context to support either long- or short-term/range communities, as summarized in Table 1.

Considering spontaneous MSNs at one extreme and online social networking services at the other, we can classify MSNs into four categories in accordance with their temporal and spatial features.

Short-Term Short-Range MSNs: They can be built on an ad hoc network that adopts wireless point-to-point communication protocols (e.g., WiFi Direct, Bluetooth), and may appear in a coffee shop, an airport departure lounge, or a moving bus. The goal is to boost face-to-face conversation or facilitate information sharing in the physical world (for both acquaintances and strangers). For instance, Meetup allows members

Digital Object Identifier: 10.1109/MCOM.2017.1700037

Zhiyong Yu is with Fuzhou University; Daqing Zhang (corresponding author) is with Telecom SudParis; Zhu Wang and Bin Guo are with Northwestern Polytechnical University; Ioanna Roussaki is with National Technical University of Athens; Kevin Doolin and Ethel Claffey are with Waterford Institute of Technology.

	Conventional MSNs	CA-MSNs
Examples	Facebook, Twitter, Google+	Foursquare, Meetup, Instagram, WeChat
Category	Mainly long-term long-range, based on Internet	All possible (detailed below), based on hybrid networks (Internet + opportunistic networks)
Context richness	No context, or only few and low-level context, such as time and location	Rich and high-level context (detailed in "Personal Context and Community Context")
Life cycle management	Manual	Automatic or aided with context (detailed in "Life Cycle Management in CA-MSNs")
Context storage	Stateless, data stored mainly on server side	Stateful, historical context stored at both client and server sides

Table 1. Comparisons of conventional and context-aware MSNs.

Context-aware computing has been recognized as a major research branch of pervasive computing since the late 1990s. Its objective was to confer more intelligence on pervasive services and systems by considering the relevant context that was not yet taken into account. Afterward, with the development of social computing, the context from social aspects drew the attention of research communities.

to create or join offline group meetings by a common temporary interest, such as books, games, movies, or pets.

Short-Term Long-Range MSNs: They usually aim to facilitate remote teamwork via the Internet to complete a large task before a given deadline, for example, voluntary support for disaster relief, like a crowdsourcing disaster support platform (CDSP) [4], which is detailed in "Our Practices on CA-MSNs."

Long-Term Long-Range MSNs: Users of social networking services extended with mobile accessibility (e.g., Facebook) form this category of MSNs. The goal is to facilitate instant messaging and information dissemination globally.

Long-Term Short-Range MSNs: They are confined to a group of people living/working together in limited physical spaces. The goal of these MSNs is to maintain relationships with familiar persons (e.g., in a family/company), with special security and privacy policies. For instance, WeChat has a function that allows users to join a private group with friends nearby.

Personal Context and Community Context

Context-Aware Computing in MSNs

Context-aware computing has been recognized as a major research branch of pervasive computing since the late 1990s [5]. Its objective was to confer more intelligence on the pervasive services and systems by considering the relevant context that was not yet taken into account. Afterward, with the development of social computing, the context from social aspects drew the attention of research communities [6, 7].

In MSNs, context refers to the information concerning not only individuals, but also multiple users and entire groups. Both *personal context* and *community context* are indispensable to intelligent decision making in each phase of the MSN life cycle. For example, if one intends to create a community by discovering nearby people with certain hobbies, it would be essential to be aware of their personal context such as interest and location. On the other hand, if one wishes to discover existing communities to join, community context such as community location and profile might be needed.

A spectrum of existing technologies has been developed for the extraction of personal and community context, including context representation, mining, and inference [8]. More noteworthy, new technologies such as community modeling, participatory sensing, and large-scale multi-modal data fusion are promising to fully empower context-aware MSNs.

Personal Context Taxonomy

Personal context describes all relevant information of a person that can characterize his/her situation. It can be classified into static personal context and dynamic personal context.

Static personal context refers to an individual profile that remains almost unchanged. It includes one's identity and affiliation, which is quite stable, and one's preference/interest, available resources, and contact list, which might change slowly. In MSNs, what we care about most is one's preferences concerning social activities, for example, likes some movies and dislikes a certain restaurant.

Dynamic personal context refers to contextual information that changes from time to time, such as one's location, physiological condition (e.g., blood pressure and heart rate), behavior (e.g., walking and laughing), activity (e.g., sleeping, meeting, in a certain mood), and intent (e.g., temporary goal for a task).

Community Context Taxonomy

Community context is able to help communities to function efficiently by exploiting and understanding the activities, similarities, and relationships of the entire community as a whole. Community context can also be static or dynamic.

Static community context consists of information about community profile and community structure. More specifically, the profile of a community includes motivation, membership, demography, resources, and preferences. The structure of a community comprises relationship (i.e., inter-personal and inter-community relationships), social status, and structural metrics obtained based on social network analysis (i.e., connection, distribution, density, and segmentation). In case members have different personal preferences, there should be a method to determine the community preference, which may be not a simple average of each member's preference.

Dynamic community context refers to the time varying contextual information of a communi-

SOCKER serves to create short-/long-term short-range MSNs. The three metrics, that is, user popularity, inter-user closeness, and user effectiveness, are used primarily for personal or community context. All can be estimated from users' historical trajectories and interactions.

	Distinctive personal context	Distinctive community context	Reference
Foursquare	Personal location	Social status	foursquare.com
Meetup	Personal preferences	Membership, community preferences	www.meetup.com
Instagram	Personal activity	Relationship, interaction	instagram.com
WeChat	Personal identity, personal activity	Community location, community motivation, interaction	www.wechat.com
SOCKER	Trajectory, personal preference	Encounter, user popularity, inter-user closeness, user effectiveness, community intent, community size	[9]
CDSP	Expertise, available time, home location	Acquaintanceship, physical proximity, interest consistency, interaction, social status, community intent	[4]
TLI	Social relationship, home location, personal preference	Overlapping influence, skill coverage, community intent, community size, activity location	[10]

Table 2. Context supported in CA-MSN products and prototypes.

ty, such as community location (e.g., proximity), intra/inter-community interaction, community activity (i.e., the abstraction of a series of interactions among community members), and community intent (i.e., the short-term common goal based on each member's requirements).

In Table 2 we report the context features supported in existing MSN products and research prototypes.

Life Cycle Management in CA-MSNs

Life Cycle of MSNs

MSNs involve the management of communities (i.e., a group of people communicating and interacting in a physical and/or virtual space for a common purpose [11]) and supporting resources (e.g., devices, networks, services). Inspired by the community management phases proposed in the EU FP7 SOCIETIES project [12], we divide the life cycle of MSNs into four phases/steps: *discover*, *connect*, *interact*, and *organize*.

Discover: discovering users, resources, services, devices, and networks for creating new communities, or discovering already existing communities for joining, merging, and splitting

Connect: connecting users to support interactions, connecting communities, or connecting members/communities to their owned devices, networks, resources and services

Interact: direct interacting via instant messaging, group chatting, and so on; indirect interacting via social media (tagging the same photos, commenting on the same videos, visiting the same places)

Organize: adding users to or removing members from communities; creating, merging, splitting, and terminating communities; managing community hierarchies, coordinating interactions among members, maintaining infrastructures of a community

Context-Aware Discovery, Connection, Interaction, and Organization

Based on the concept of context and life cycle, the methodology can be described as: in each life cycle phase, diverse context should be exploited to make MSNs more intelligent, as shown in Fig. 1.

Context-Aware Discovery: In order to create a new community or identify an existing community to join, the first step is to discover the related people and resources crossing the boundary between the physical and virtual space. We note that current systems allow to some extent the discovery of people and devices in the physical environment via the Internet of Things (IoT, e.g., RFID tags) or

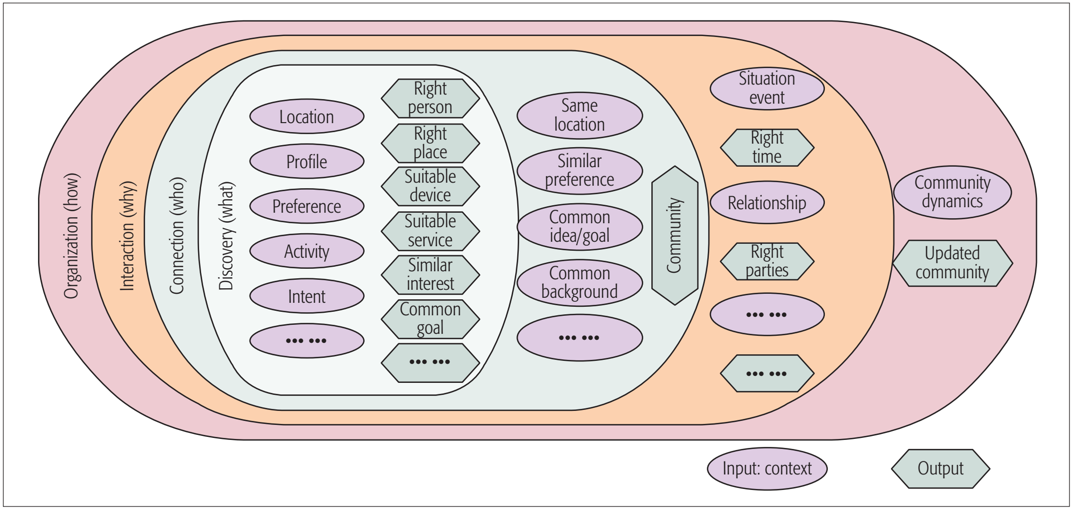


Figure 1. Context-aware discovery, connection, interaction, and organization.

in cyberspace via searching on the Internet. However, these systems do not thoroughly exploit the variety of context for socializing purposes. As a an MSN aims to pull people and resources together, the most important context should include personal/community location, preference, intent, activity, and so on.

Context-Aware Connection: People can possibly choose a communication channel from a wide range of methods (e.g., Internet or ad hoc network, text, or video). Personal and community context can ease a series of issues like connection establishment and switching, and can help to choose the most efficient connection from a quality and cost perspective. For example, a group chat is taking place locally, and the members are sharing a video via a mobile ad hoc network; then a remote friend wants to join the chat, and the ad hoc network can connect with the Internet automatically to enable the remote friend to receive the video.

Context-Aware Interaction: Both the community context and members' personal context play a vital role in enabling humanistic social interaction. A major challenge here is to identify the events and situations that should trigger the interactions. Such events and situations might vary greatly. A simple event may involve two members being available for a chat, while a complex situation could be friends negotiating a local tour based on their interests and free time. In general, the relationship and commonality between interaction parties should be monitored.

Context-Aware Organization: This task includes introducing new members to an established community through additional discovery/connection steps, and removing members who are not relevant to the community anymore. A major challenge here is to also detect events and situations that would trigger community evolution, such as joining, leaving, splitting, merging, and so on. Community dynamics such as change of location, interaction, activity, and intent can be exploited to support adaptive membership management.

Our Practices on CA-MSNs

SOCKER: Socially Aware Broker-Based Community Creation Mechanism

Various methods can rally people for a local activity, for example, posting a public announcement saying that there will be a weekend party at a nearby bar. However, some extra concerns include:

- The number of attendees should be controlled precisely, but not in a competitive way.
- The information of the activity, its attendees, and the sifting process should be kept private.
- The initiator of the activity has certain social expectations, for example, to make new friends or to entertain old friends.

Overlooking these concerns would disappoint rejected individuals. Our work, SOCKER [9], a socially aware broker-based community creation mechanism, aims to get together like-minded persons for a particular face-to-face social activity, with consideration of the above concerns.

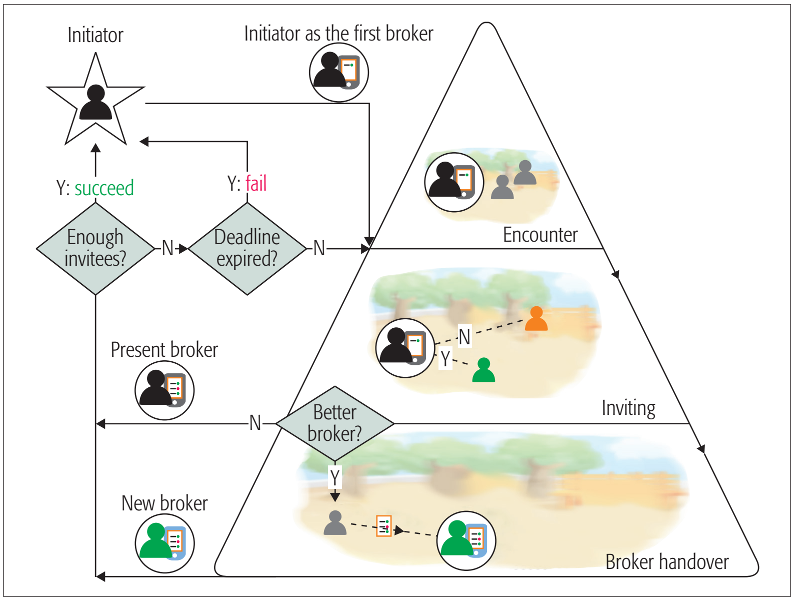
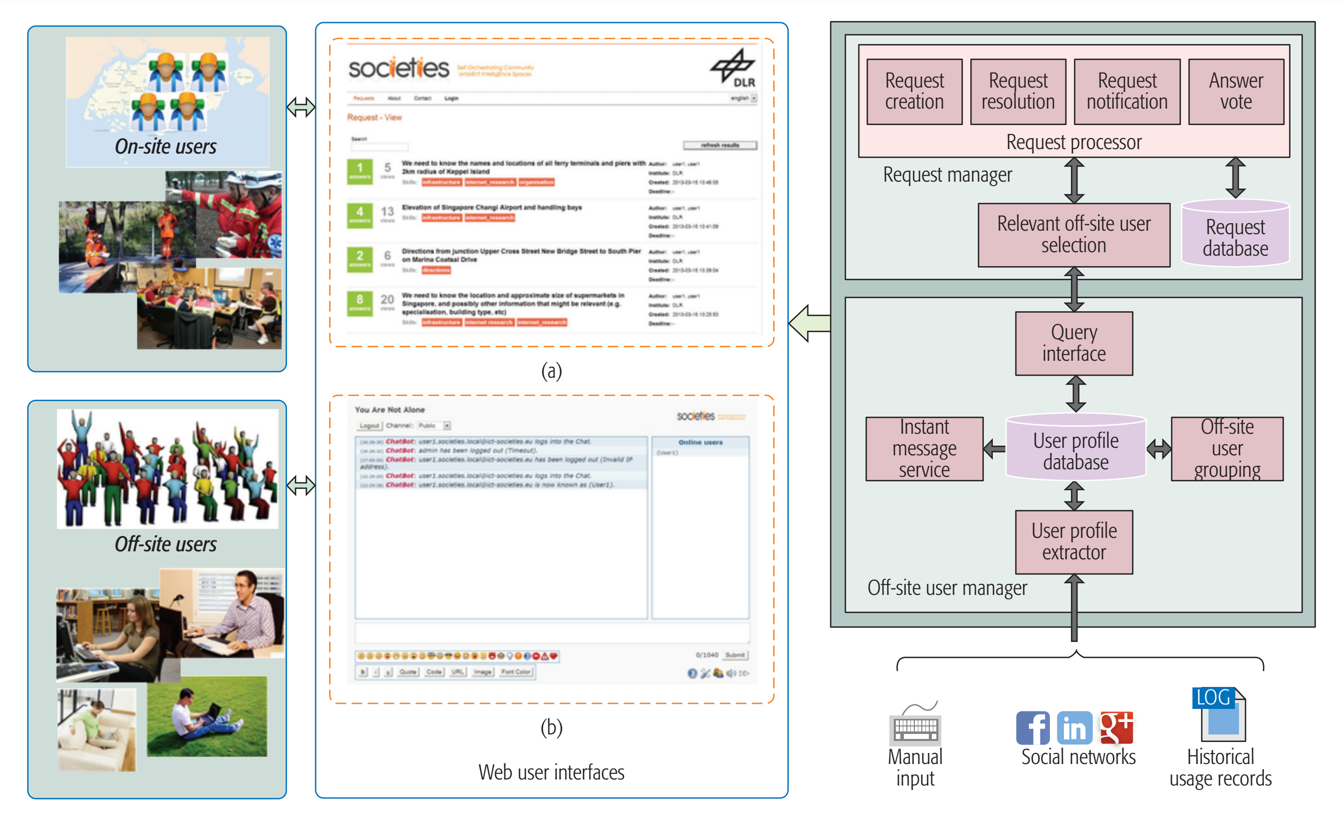


Figure 2. Community creation procedure of SOCKER.

SOCKER regards the community creation problem as a broker-based information dissemination task. Three metrics (detailed below) are measured to judge whether a person is a proper broker. Figure 2 illustrates the procedure of SOCKER. Concretely, the initiator serves as the first broker when he plans to create a new community. As the broker moves in the physical world, he/she will encounter other persons opportunistically. For each person u_i she/he meets, the broker decides whether to invite u_i to join the community (according to social expectations of the initiator and matchmaking between the activity type and u_i's preference) and updates progress records such as current invitees and met-but-uninvited user list. Meanwhile, SOCKER decides whether to hand over the broker role to u_i. If the broker handover condition is satisfied according to the three predefined metrics, the current broker will send the progress records to the new broker, then stop acting as a broker. The new broker will continue the community creation task just as her/his predecessor was doing it. The task is completed successfully if the required community size is reached before the deadline. If so, the last broker will report on the task accomplishment to all the invitees and the activity initiator. Otherwise, when the deadline expires but the current community size is still smaller than needed, the last broker will notify the activity initiator saying that the community creation task has failed.

SOCKER serves to create short-/long-term short-range MSNs. The three metrics, that is, user popularity, inter-user closeness, and user effectiveness, are used primarily for personal or community context. All can be estimated from users' historical trajectories and interactions. For a specific user, the user popularity is defined as the number of different persons she/he will encounter in a forthcoming period (e.g., a week). Intuitively, a user with higher popularity tends to meet more people, and thereby community creation can be accelerated. For two users, their inter-user closeness is defined as the number of encounters



1 answers 5 views — We need to know the names and locations of all ferry terminals and piers with 2km radius of Keppel Island — Skills: infrastructure, internet_research, organisation — Author: user1, user1 — Institute: DLR — Created: 2013-03-15 13:46:05 — Deadline: -

4 answers 13 views — Elevation of Singapore Changi Airport and handling bays — Skills: infrastructure, internet_research — Author: user1, user1 — Institute: DLR — Created: 2013-03-15 13:41:09 — Deadline: -

2 answers 6 views — Directions from junction Upper Cross Street New Bridge Street to South Pier on Marina Coatsal Drive — Skills: directions — Author: user1, user1 — Institute: DLR — Created: 2013-03-15 13:39:04 — Deadline: -

8 answers 20 views — We need to know the location and approximate size of supermarkets in Singapore, and possibly other information that might be relevant (e.g. specialisation, building type, etc) — Skills: infrastructure, internet research, internet_research — Author: user1, user1 — Institute: DLR — Created: 2013-03-15 13:28:53 — Deadline: -

ChatBot: user1.societies.local@ict-societies.eu logs into the Chat.
ChatBot: admin has been logged out (Timeout).
ChatBot: user1.societies.local@ict-societies.eu has been logged out (Invalid IP address).
ChatBot: user1.societies.local@ict-societies.eu logs into the Chat.
ChatBot: user1.societies.local@ict-societies.eu is now known as (User1).

Figure 3. System overview of CDSP.

in a forthcoming period. The inter-user closeness of two successive brokers should be higher than a threshold when the activity initiator wants to play with old friends, while vice versa to make new friends. User effectiveness considers the progress records obtained during the community growing process to avoid broker handover in this case: although the candidate new broker would encounter many users in the future, perhaps previous brokers have already encountered most of them. It is defined as the number of unrecorded users one will encounter in a forthcoming period.

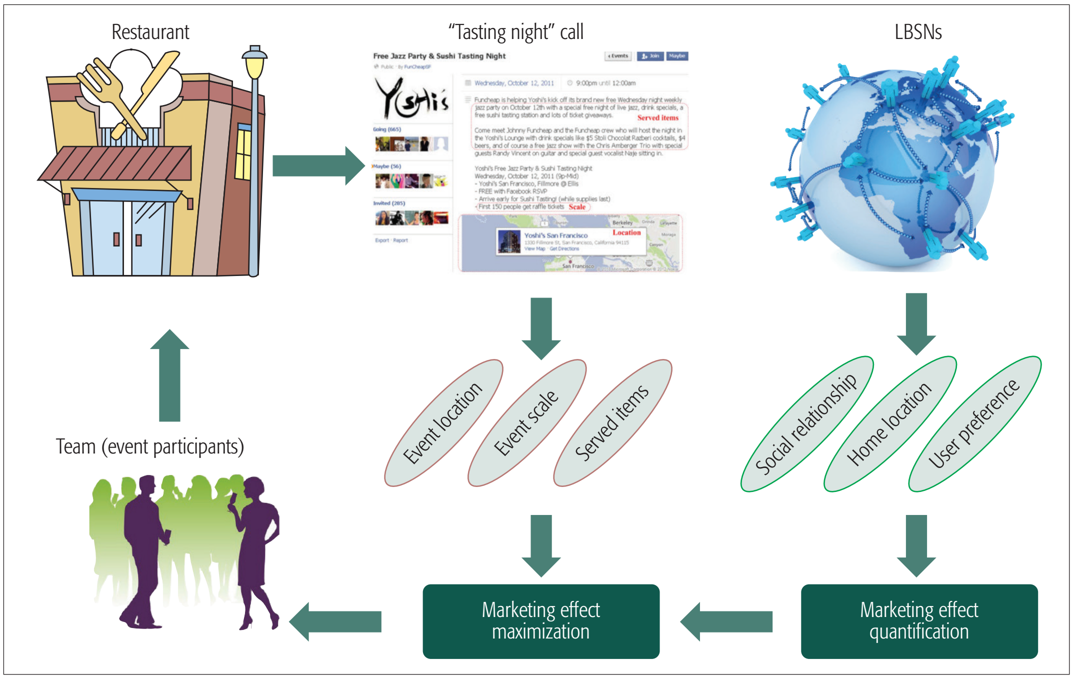
Our experience shows that more communities can be created successfully with lower costs when with the help of selected context. User popularity is helpful to meet enough users rapidly, inter-user closeness can increase the chance of encountering the right persons, and user effectiveness is able to minimize the risk of appointing inappropriate brokers.

Crowdsourcing Disaster Support Platform

When a disaster (e.g., an earthquake) occurs, we need a support system to help rescue teams in saving lives, property, and the environment. On-site rescue teams may meet with many problems such as finding a passable road or recognizing a nameplate written in an unknown language. Crowdsourcing is a feasible way to subcontract these tasks/requests to a large number of off-site volunteers. Therefore, CDSP [4] was developed in the SOCIETIES project. With CDSP, volunteers from all over the world can share the burden of large tasks, take turns responding to requests immediately, and interact with each other to make the answers more credible.

In order to avoid presenting a long request list to bother off-site users (i.e., volunteers), we introduce the skill-matching mechanism. On one hand, performing a task may involve some kinds of expertise, which is specified by the request creator (e.g., an on-site user). For example, damage assessment from satellite images needs the skill of image analysis. On the other hand, CDSP can discover users' expertise from various sources, such as social networks, historical usage records of the platform, and manual input. The matching mechanism makes sure that the tasks are assigned to those who can handle them, and irrelevant requests are screened. For a large task, users first divide it into sub-tasks, and then work on these sub-tasks in parallel, and finally merge and output answers. Furthermore, they can have conversations through an instant messaging interface. Figure 3 shows the system overview of CDSP.

The "off-site user grouping" component retrieves a group of users whose expertise is demanded by the large task, then connects them with a short-term long-range MSN, whose goal is to facilitate information sharing or collaborative work to accomplish a disaster relief task. To manage the MSN life cycle, both personal and community *context* are useful. Besides expertise, other personal context considered includes available time and home location. For example, a plenary meeting needs all members to be available at the same time, while monitoring a webcam incessantly needs their available times should not be necessary. Some tasks are more favorable to users from a particular regional background (i.e., home location). For instance, nameplate recognition is easier for users from regions near the disaster site. The above is for the discovery and connection phase. For the interaction and organization phase, com-



Free Jazz Party & Sushi Tasting Night

Public · By FunCheapSF

Wednesday, October 12, 2011 · 9:00pm until 12:00am

Funcheap is helping Yoshi's kick off its brand new free Wednesday night weekly jazz party on October 12th with a special free night of live jazz, drink specials, a free sushi tasting station and lots of ticket giveaways.

Come meet Johnny Funcheap and the Funcheap crew who will host the night in the Yoshi's Lounge with drink specials like $5 Stoli Chocolat Razberi cocktails, $4 beers, and of course a free jazz show with the Chris Amberger Trio with special guests Randy Vincent on guitar and special guest vocalist Naje sitting in.

Yoshi's Free Jazz Party & Sushi Tasting Night
Wednesday, October 12, 2011 (9p-Mid)
- Yoshi's San Francisco, Fillmore @ Ellis
- FREE with Facebook RSVP
- Arrive early for Sushi Tasting! (while supplies last)
- First 150 people get raffle tickets

Yoshi's San Francisco
1330 Fillmore St, San Francisco, California 94115
View Map · Get Directions

Figure 4. Team formation procedure of TLI.

Lessons from this practice include: personal context is useful for specifying and testing the condition to join a community; members' interaction can help to complete a large task more effectively and efficiently; and we can rely on community context from multiple sources to boost the interaction.

munity context is newly gathered or inferred from personal context. Acquaintanceship (e.g., linkage on social networks), physical proximity (e.g., from the same region), and interest consistency (e.g., with many overlapped skills) can be leveraged to warm up and boost the interaction among community members. Typical interactions include brainstorming and voting, which can improve the answer's credibility. When a task is finished, the community intent does not exist, and all members go back to the pool of volunteers, waiting/looking for the next task.

Lessons from this practice include: personal context is useful for specifying and testing the condition to join a community; members' interaction can help to complete a large task more effectively and efficiently; and we can rely on community context from multiple sources to boost the interaction.

TLI: Team Formation with Local Influence for Offline Event Marketing

Offline events are favorable ways for business owners to transform current customers into brand advocates and reach potential customers. For example, a restaurant hosts a tasting night, invites a certain number of participants, and hopes that after the event these participants can attract more users (by their social influence) to visit this restaurant. Who should be invited? This is a problem of influence maximization with limited budget. However, several factors should be considered in this specific application. Unlike existing works that simply count how many users a single celebrity influences, we depict influence by estimating p_i, which denotes the possibility of a user to visit a venue (viz. where the business is located, e.g., a restaurant). First, p_i depends on how many influencers (i.e., event participants) have exerted influence on the user. A team's influence on a user is not straightforwardly the sum of individual members' influence, but follows the law of diminishing marginal utility. We name it the *overlapping* factor. Second, p_i depends on the distance between the user and the venue, the so called *distance* factor. Third, p_i depends on whether the influencer and the user like the brand, specifically, the products or services offered by the business, named the *coverage* factor. In order to tackle these factors, we propose TLI [10], a team formation approach with local influence.

We first build a marketing effect quantitative model with considerations of the factors of overlapping, distance, and coverage. Then a combinatorial optimization problem is formulized for the marketing effect maximization, which is approximately solved with a heuristic algorithm. The participant team can be determined accordingly. Figure 4 shows the team formation procedure of TLI.

Apparently, team formation is a special type of community creation. The team members should meet certain constraints (e.g., skill coverage) while maximizing another metric (e.g., working achievement). This makes a team not as flexible or dynamic as an ordinary community, that is, members cannot join or leave a team at any time. In order to form a (near) optimal team, several kinds of personal and community *context* are collected and analyzed comprehensively, including social relationship, home location, and user preference, from location-based social networks. Users' social relationships are used to eliminate overlapping social influence. Users' home locations are used for discovering the relationship between the visiting probabilities and the distances. Since previous works prove that it follows the power law, we train the parameters of a power function from check-in records. Users' preferences are extracted from visiting histories and joined together to ensure that the participant team covers all served products. Based on the above context awareness, we build an accurate and fine-grained influence model for a team.

Until now, MSNs have revolutionized the style people interact and communicate. The next generation of MSNs is expected to not only facilitate interaction and communication among people with better effectiveness, but also match the demand and the supply (in terms of information, services and goods) among people in a more intelligent manner.

We learned that when simple context cannot meet the requirements of an application, we need to design even richer and higher-semantic-level context, which can be inferred from low-level context with the help of technologies including community modeling, context mining, and so on.

Conclusion and Future Directions

Until now, MSNs have revolutionized the style in which people interact and communicate. The next generation of MSNs is expected to not only *facilitate interaction and communication* among people with better effectiveness, but also *match the demand and supply* (in terms of information, services, and goods) among people in a more intelligent manner [13]. This is particularly pertinent to the many businesses that desire a strong online social presence, but remain challenged by this concept [14]. The CA-MSN is a promising evolution direction of MSNs; they could be enhanced from the following perspectives.

Extending the Sensing Capability of MSNs with IoT and Mobile Crowd Sensing: By leveraging the sensors embedded in mobile devices, sensor networks installed in our surroundings, and the human digital footprints recorded by the IoT, a huge amount of context about users and their interactions in MSNs can be acquired. With these context data and corresponding big data technologies, more intelligent matchmaking and interaction among mobile users and resources can be supported. For example, by collecting bus/metro card records for a period, the system can learn members' daily movement patterns, and then the community's planning the location and time of a gathering can be more convenient. Mobile crowd sensing utilizes citizens' off-the-shelf smartphones to capture social and urban dynamics [15]. By leveraging human power in the loop of the sensing and computing process, MSNs have the most favorable position to gain the advantages of the crowdsourced context.

Extending the Communication Capability of MSNs by Bridging Mobile Ad Hoc Networks to Infrastructure-Based Networks Seamlessly: Nowadays, most existing online social networking services lack effective support for face-to-face interaction in the physical world, especially when/where no infrastructure is available. This fact calls for research on the creation, organization, and migration of offline social networks to online social networks, and seamless transition between online and offline social networks. That is to say, future MSNs should be infrastructure-independent and capable of supporting both long-term relationships and spontaneous social interactions. For example, a user meeting others opportunistically in a coffee shop can create a local community via a mobile ad hoc network. When she leaves the coffee shop, this community can still be maintained online for further interactions.

Extending the Service Platform of MSNs with Context Awareness Features: Generally speaking, MSNs can be seen as a service platform to ease information sharing, user interactions, service discovery/consumption, and users' personal demand/satisfaction. To provide a convenient and effective service platform for mobile users, it is advisable to enhance key features of the platform with context awareness. Last but not least, the MSN should be an open platform that enables developers to freely create new context-aware applications for specific needs.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China (nos. 61772136, 61772428, 61402369, 61373119, 61672159), the Fujian Collaborative Innovation Center for Big Data Application in Governments, and the Technology Innovation Platform Project of Fujian Province (no. 2014H2005).

References

[1] N. Vastardis and Y. Kun, "Mobile Social Networks: Architectures, Social Properties, and Key Research Challenges," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 3, 2013, pp. 1355–71.

[2] D. Zhang *et al.*, "The Emergence of Social and Community Intelligence," *IEEE Computer*, vol. 44, no. 7, 2010, pp. 21–28.

[3] H. Chen *et al.*, "A Generic Framework for Constraint-Driven Data Selection in Mobile Crowd Photographing," *IEEE Internet of Things J.*, vol. 4, no. 1, 2017, pp. 284–96.

[4] D. Yang *et al.*, "Providing Real-Time Assistance in Disaster Relief by Leveraging Crowdsourcing Power," *Personal and Ubiquitous Computing*, vol. 18, no. 8, 2014, pp. 2025–34.

[5] A. Dey *et al.*, "A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications," *HumanComputer Interaction*, vol. 16, no. 2, Dec. 2001, pp. 97–166.

[6] Z. Yu *et al.*, "Personalized Travel Package with Multi-Point-of-Interest Recommendation Based on Crowdsourced User Footprints," *IEEE Trans. Human-Machine Systems*, vol. 46, no. 1, 2016, pp. 151–58.

[7] A. Pentland, "Socially Aware, Computation and Communication," *IEEE Computer*, vol. 38, no. 3, 2005, pp. 33–40.

[8] X. Wang *et al.*, "Ontology-Based Context Modeling and Reasoning Using OWL," *Proc. PerCom Wksp.*, Orlando, FL, Mar. 2004, pp. 18–22.

[9] Z. Wang *et al.*, "SOCKER: Enhancing Face-to-Face Social Interaction Based on Community Creation in Opportunistic Mobile Social Networks," *Wireless Personal Commun.*, vol. 78, no. 1, 2014, pp. 755–83.

[10] Z. Yu *et al.*, "Participant Selection for Offline Event Marketing Leveraging Location-Based Social Networks," *IEEE Trans. Systems, Man, Cybernetics: Systems*, vo. 45, no. 6, 2015, pp. 853–64.

[11] N. Lane, "Community-Aware Smartphone Sensing Systems," *IEEE Internet Computing*, vol. 16, no. 3, 2012, pp. 60–64.

[12] I. Roussaki *et al.*, "Context Awareness in Wireless and Mobile Computing Revisited to Embrace Social Networking," *IEEE Commun. Mag.*, vol. 50, no. 6, June 2012, pp. 74–81.

[13] A. Chin, "Ephemeral Social Networks," *Mobile Social Networking: An Innovative Approach*, Springer, 2013, pp 25–64.

[14] E. Claffey and M. Brady, "A Model of Consumer Engagement in a Virtual Customer Environment," *J. Customer Behaviour*, vol. 13, no. 4, 2014, pp. 325–46.

[15] B. Guo *et al.*, "ActiveCrowd: A Framework for Optimized Multi-Task Allocation in Mobile Crowdsensing Systems," *IEEE Trans. Human-Machine Systems*, vol. 47, no. 3, 2017, pp. 392–403.

Biographies

Zhiyong Yu (yuzhiyong@fzu.edu.cn) is an associate professor at the College of Mathematics and Computer Science, Fuzhou University, China, also affiliated with the Fujian Provincial Key Laboratory of Network Computing and Intelligent Information Processing and the Key Laboratory of Spatial Data Mining and Information Sharing, Ministry of Education, China. He received his Ph.D. from Northwestern Polytechnical University, China, in 2011. He was a visiting student at Kyoto University, Japan, from 2007 to 2009 and a visiting researcher at Telecom SudParis, France, from 2012 to 2013. His current research interests include pervasive computing, mobile social networks, and crowdsensing.

Daqing Zhang (daqing.zhang@telecom-sudparis.eu) is a professor with Telecom SudParis and SAMOVAR, CNRS, France. He obtained his Ph.D. from the University of Rome "La Sapienza," Italy, in 1996. His research interests include context-aware computing, urban computing, mobile computing, and so on. He has served as the General or Program Chair for more than 10 inter-

national conferences. He is an Associate Editor of *ACM Transactions on Intelligent Systems and Technology*, *IEEE Transactions on Big Data*, and other publications.

Zhu Wang (wangzhu@nwpu.edu.cn) is an associate professor at the School of Computer Science, Northwestern Polytechnical University. He received his Ph.D. in computer science and technology from Northwestern Polytechnical University in 2013. from November 2010 to April 2012, he worked as a visiting student at Telecom SudParis. His research interests include pervasive computing, mobile social networking, and healthcare.

Bin Guo (guob@nwpu.edu.cn) is a professor at the School of Computer Science, Northwestern Polytechnical University. He received his Ph.D. from Keio University, Tokyo, Japan, in 2009. During 2009–2011, he was a postdoctoral researcher at Telecom SudParis. His research interests include pervasive computing, social computing, and mobile crowdsensing. He has served as an Editor or Guest Editor for a number of international journals, such as *IEEE THMS* and *ACM TIST*.

Ioanna Roussaki (ioanna.roussaki@cn.ntua.gr) received her Diploma in electrical and computer engineering in 1997 from the National Technical University of Athens (NTUA), Greece. In 2003, she received her Ph.D. in the area of telecommunications and computer networks. She has participated in many national and international research and development projects. Since 2015, she has been an assistant professor in the NTUA School of Electrical and Computer Engineering. Her research interests include the Internet of Things, context awareness, social computing, and so on.

Kevin Doolin (kdoolin@tssg.org) is director of Innovation at Waterford Institute of Technology's Telecommunications Software and Systems Group (TSSG). His area of expertise is pervasive computing, which is the forerunner to the Internet of Things. He has coordinated a number of key EU projects in this space, including PERSIST (www.ict-persist.eu) and SOCIETIES (www.ict-societies.eu), which closed in April 2014 and received significant praise from expert scientific reviewers, and which focused on the integration of pervasive and social computing.

Ethel Claffey (ECLAFFEY@wit.ie) is a lecturer in marketing in theSchool of Business at Waterford Institute of Technology. Her Ph.D. was awarded by Trinity College Dublin. Her research interests include consumer engagement, contemporary consumer behavior, virtual communities, technology acceptance, and digital marketing. Her work has been published in a variety of conference proceedings and refereed journal articles such as *Psychology & Marketing* and the *Journal of Customer Behaviour*.

Non-Orthogonal Multiple Access in Multi-Cell Networks: Theory, Performance, and Practical Challenges

Wonjae Shin, Mojtaba Vaezi, Byungju Lee, David J. Love, Jungwoo Lee, and H. Vincent Poor

The authors discuss the opportunities and challenges of NOMA in a multi-cell environment. As the density of base stations and devices increases, inter-cell interference becomes a major obstacle in multi-cell networks. As such, identifying techniques that combine interference management approaches with NOMA is of great significance.

Abstract

Non-orthogonal multiple access (NOMA) is a potential enabler for the development of 5G and beyond wireless networks. By allowing multiple users to share the same time and frequency, NOMA can scale up the number of served users, increase spectral efficiency, and improve user-fairness compared to existing orthogonal multiple access (OMA) techniques. While single-cell NOMA has drawn significant attention recently, much less attention has been given to multi-cell NOMA. This article discusses the opportunities and challenges of NOMA in a multi-cell environment. As the density of base stations and devices increases, inter-cell interference becomes a major obstacle in multi-cell networks. As such, identifying techniques that combine interference management approaches with NOMA is of great significance. After discussing the theory behind NOMA, this article provides an overview of the current literature and discusses key implementation and research challenges, with an emphasis on multi-cell NOMA.

What Drives NOMA?

The next generation of wireless networks will require a paradigm shift in order to support massive numbers of devices with diverse data rates and latency requirements. In particular, the increasing demand for Internet of Things (IoT) devices poses challenging requirements on 5G wireless systems. Two key features of 5G are expected to be a latency of 1ms, compared to 10 ms in the 3rd Generation Partnership Project (3GPP) Long-Term Evolution (LTE), and support for 10 Gb/s throughput.

To fulfill these requirements, numerous potential technologies have been introduced over the last few years. Among them is non-orthogonal multiple access (NOMA) [1], a technique to serve multiple users via a single wireless resource. NOMA can be realized in the power, code, or other domains [2, 3]. Code domain NOMA uses user-specific spreading sequences for sharing the entire resource, whereas power domain NOMA exploits the channel gain differences between the users for multiplexing via power allocation.

Power domain NOMA can improve wireless communication with the following benefits.

Massive Connectivity: There appears to be a reasonable consensus that NOMA is essential for massive connectivity, because the number of served users in all orthogonal multiple access (OMA) techniques is inherently limited by the number of resource blocks. In contrast, NOMA theoretically can serve many users in each resource block by superimposing the users' signals. In this sense, NOMA can be tailored to typical IoT applications where a large number of devices sporadically try to transmit small packets.

Low Latency: Latency requirements for 5G applications are rather diverse. Unfortunately, OMA cannot guarantee such broad delay requirements because no matter how many bits a device wants to transmit, the device must wait until an unoccupied resource block is available. In contrast, NOMA supports flexible scheduling since it can accommodate a *variable* number of devices depending on the application that is being used and the perceived quality of service (QoS) of the device.

High Spectral Efficiency: NOMA also surpasses OMA in terms of spectral efficiency and user-fairness. As will be seen later, NOMA is a theoretically optimal way of using spectrum for both uplink and downlink communications in a single-cell network. This is because every NOMA user can enjoy the whole bandwidth, whereas OMA users are limited to a smaller fraction of spectrum which is inversely proportional to the number of users. In addition, NOMA can also be combined with other emerging technologies, such as massive multiple-input multiple-output (MIMO) and millimeter wave (mmWave) technologies, to further support higher throughput.

In view of the above benefits, NOMA has drawn much attention from both academia and industry. However, much of the work in this context is limited to single-cell analysis, where there is no co-channel interference caused by an adjacent base station (BS). To verify the benefits of NOMA in a more realistic setting, it is necessary to consider a multi-cell network. Specifically, as wireless networks become denser and denser, inter-cell interference (ICI) becomes a major obstacle to achieving the benefits of NOMA. In this regard, we consider NOMA in a multi-cell environment for this article. We first discuss the theory behind NOMA and an overview of the literature on

Digital Object Identifier:
10.1109/MCOM.2017.1601065

Wonjae Shin and Jungwoo Lee are with Seoul National University; Mojtaba Vaezi and H. Vincent Poor are with Princeton University; Byungju Lee (corresponding author) and David J. Love are with Purdue University.

NOMA. We then explain the main implementation issues and research challenges, with a particular focus on multi-cell NOMA. Finally, the system-level performance evaluation of multi-cell NOMA solutions will be provided before concluding the article.

Theory Behind NOMA

Analysis of cellular communication can generally be classified as either *downlink* or *uplink*. In the downlink channel, the BS simultaneously transmits signals to multiple users, whereas in the uplink channel multiple users transmit data to the same BS. From an information-theoretic perspective, the downlink and uplink are modeled by the *broadcast channel* (BC) and *multiple access channel* (MAC), respectively. The basic premise behind single-cell NOMA in the power domain is to reap the benefits promised by the theory of multi-user channels [4]. As such, we review what information theory promises for these channels, both in the single-cell and multi-cell settings. In particular, we seek to answer the following two questions in this section:

- What are the highest achievable throughputs for these multi-user channels?
- How can a system achieve such rates?

Single-Cell NOMA

The capacity regions of the two-user MAC and BC are achieved via NOMA, where both users' signals are transmitted at the same time and in the same frequency band [5]. The curves labeled by NOMA in Figs. 1a and 1b represent the MAC and BC capacity regions, respectively, for the case of additive white Gaussian noise (AWGN). Except for a few points, OMA is strictly suboptimal, as can be seen from the figures. To gain more insight, we describe how the above regions are obtained. For OMA we consider a time division multiple access (TDMA) technique where α fraction of time ($0 \leq \alpha \leq 1$) is dedicated to user 1 and $\bar{\alpha} \triangleq 1-\alpha$ fraction of time is dedicated to user 2. In this article, we use the following notation:

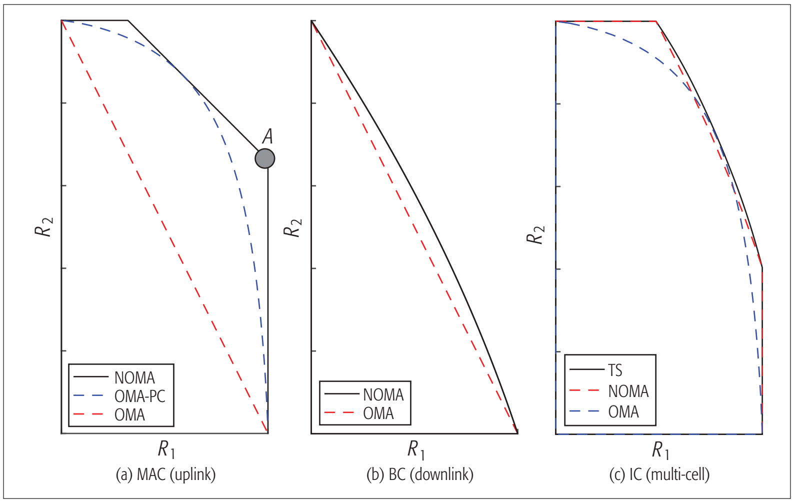
$$\mathcal{C}(x) \triangleq \frac{1}{2}\log_2(1+x)$$

and $\gamma_i = |h_i|^2 P$ is the received signal-to-noise ratio (SNR) for user i, where h_i is the channel gain, P is the transmitter power, and the noise power is normalized to unity.

Uplink (MAC): Using OMA, each user sees a single-user channel in its dedicated fraction of time, and thus $R_1 = \alpha\mathcal{C}(\gamma_1)$ and $R_2 = \bar{\alpha}\mathcal{C}(\gamma_2)$ are achievable. If power control is applied, these rates can be boosted to $R_1 = \alpha\mathcal{C}(\gamma_1/\alpha)$ and $R_2 = \bar{\alpha}\mathcal{C}(\gamma_2/\bar{\alpha})$. In the case of NOMA, both users concurrently transmit, and their signals interfere with each other at the BS. The BS can use *successive interference cancellation* (SIC) to achieve any point in the NOMA region, which is the capacity region of this channel [4]. In particular, to achieve point A the BS first decodes user 2's signal treating the other signal as noise. This results in

$$R_2 = \mathcal{C}\left(\frac{\gamma_2}{\gamma_1+1}\right).$$

The BS then removes user 2's signal and decodes user 1's signal free of interference, that is, $R_1 = \mathcal{C}(\gamma_1)$. From Fig. 1a it is seen that the gap between the NOMA and OMA regions becomes larger if power control is not used in OMA.



(a) MAC (uplink) (b) BC (downlink) (c) IC (multi-cell)

Figure 1. Best achievable regions by OMA and NOMA in the multiple access channel (MAC), broadcast channel (BC), and interference channel (IC).

Downlink (BC): In the downlink, OMA can only achieve $R_1 = \alpha\mathcal{C}(\gamma_1)$ and $R_2 = \bar{\alpha}\mathcal{C}(\gamma_2)$. However, making use of a NOMA scheme can strictly increase this rate region as shown in Fig. 1b. In particular, the capacity region of this channel is known and can be achieved using *superposition coding* at the BS. For decoding, the user with the stronger channel uses SIC to decode its signal free of interference, that is, $R_1 = \mathcal{C}(\beta\gamma_1)$, while the other user is capable of decoding at a rate of

$$R_2 = \mathcal{C}\left(\frac{\bar{\beta}\gamma_2}{\beta\gamma_2+1}\right),$$

where β is the fraction of the BS power allocated to user 1's data and $\bar{\beta} = 1 - \beta$. By varying β from 0 to 1, any rate pair (R_1, R_2) on the boundary of the capacity region of the BC (NOMA region) can be achieved.

The fact that the capacity region of downlink NOMA is known enables us to find the optimum power allocation corresponding to any point (R_1, R_2) on the boundary of the capacity region. In fact, all we need to know to achieve such a rate pair is to find what fraction of the BS power should be allocated to each user. Corresponding to each (R_1, R_2) there is a $0 \leq \beta \leq 1$ such that βP and $\bar{\beta}P$ are the optimal powers for user 1 and user 2, respectively, where P is the BS power. Conversely, every β generates a point on the boundary of the capacity region.

The above argument implies that NOMA can improve *user-fairness* smoothly and optimally by flexible power allocation. Suppose that a user has a weak channel. To boost this user's rate and improve user-fairness the BS can simply increase the fraction of power allocated to this user. We can look at this problem from yet another perspective. To increase the rate of such a user, we can maximize the weighted sum-rate $\mu R_1 + R_2$ where a high weight (μ) is given to such a user. This is because, to maximize $\mu R_1 + R_2$ for any $\mu \geq 0$ there exists an optimal power allocation strategy, determined by β. Seeing that $\mu > 1$ ($\mu < 1$) corresponds to the case where user 1 has a higher (lower) weight than user 2, to improve the user-fairness we can assign an appropriate weight to the important user and find the corresponding β.

Optimal uplink and downlink transmit/receive strategies for multi-cell networks is unknown in general. However, a combination of NOMA and OMA, known as the Han-Kobayashi scheme, gives a larger achievable region compared to the cases in which OMA or NOMA are used single-handedly.

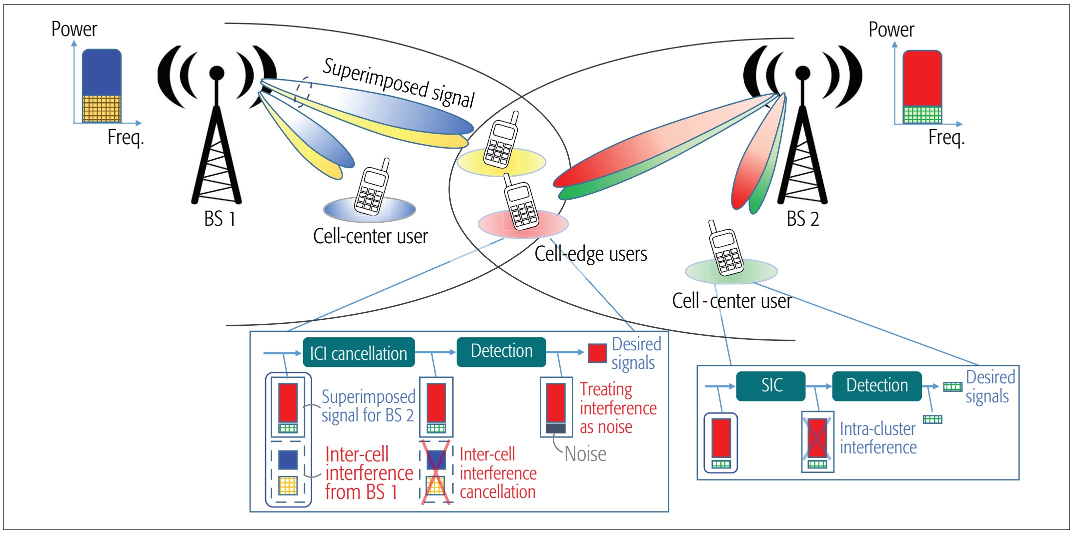


Figure 2. An illustration of multi-cell NOMA networks.

***K*-User Uplink/Downlink:** In the above, we described coding strategies for the two-user uplink/downlink channels. Interestingly, very similar coding schemes are still capacity-achieving for the *K*-user MAC and BC, that is, superposition coding with SIC gives the largest region for the *K*-user BC. Similarly, to achieve the capacity region of the *K*-user MAC, the users transmit their signals concurrently and the BS applies SIC, as described in [4, Section 6.1.4]. These schemes are based on NOMA as they allow multiple users to transmit at the same time and frequency. Additionally, OMA is strictly suboptimal [4].

Multi-Cell NOMA

In a multi-cell setting, these problems are more involved and simple channel models are insufficient. Unfortunately, capacity-achieving schemes are unknown. However, the achievable rate regions for the interference channel indicate suboptimality of OMA, as shown in Fig. 1c.

Interference Channel (IC): The capacity region of the two-user IC is not known in general; however, it is known that OMA is strictly sub-optimal. The Han-Kobayashi (HK) scheme [5] is the best known achievable scheme for the IC. In its basic form, the HK scheme employs *rate-splitting* and *superposition coding* at each transmitter. Since it uses superposition coding, the basic HK makes use of a NOMA. In general, the HK scheme applies *time-sharing* to improve the basic HK region and can be seen as a combination of NOMA and OMA [6].

The HK scheme that combines NOMA and OMA gives the largest rate region [6], as shown in Fig. 1c. In this figure, OMA refers to TDMA, whereas NOMA refers to the basic HK scheme in which time-sharing is not applied. The third curve, labeled TS, is based on the HK scheme with time-sharing (TS) in which two time slots are used. In one time slot both users are active while in the other time slot only one of them is transmitting. As can be seen from this figure, both NOMA and OMA are suboptimal when compared with the case where NOMA and OMA are combined.

Interfering MAC and BC: Consider a mutually interfering two-cell network in the uplink, where each cell includes one MAC. Assume that only one of the transmitters of each MAC (typically the closest one to the cell-edge) is interfering with the BS of the other MAC. In this network, the interfering transmitters can employ HK coding, similar to that used in the IC, while the non-interfering transmitters in each MAC employ single-user coding. This NOMA-based transmission results in an inner bound that is within a one-bit gap of the capacity region [7]. Likewise, one can use an interfering BC to model a mutually interfering two-cell downlink network.

Despite years of intensive research, finding optimal uplink and downlink transmit/receive strategies for multi-cell networks remains rather elusive. In fact, as discussed earlier, even for a much simpler case of the two-user IC, the optimal coding strategy is still unknown. Nonetheless, fundamental results from information theory as a whole suggest that NOMA-based techniques result in a superior rate region when compared with OMA.

It should be highlighted that despite the above insight from information theory, OMA techniques have been used in the cellular networks from 1G to 4G, mainly to avoid interference.[1] In addition, the lack of understanding of optimal strategies for multi-cell networks has motivated pragmatic approaches in which interference is simply treated as noise.

Single-Cell NOMA: A Review

As explained earlier, the basic theory of NOMA has been around for several decades. However, a new wave of research on NOMA has been motivated by the advance of processors that makes it possible to implement SIC at the user equipment. Saito *et al*. [1] first observed the potential of NOMA for 5G systems. They showed that NOMA can improve system throughput and user-fairness over orthogonal frequency division multiple access (OFDMA). Since then, NOMA has attracted considerable attention from both industry and academia. To make this concept more practical, several issues such as user-pairing, power allocation, and SIC implementation issues have been studied in [8]. NOMA has also been considered in the 3GPP

[1] For 3G, wideband code-division multiple access (WCDMA) was adopted, wherein *orthogonal* channelization codes are used within a cell, yet *quasi-orthogonal* scrambling codes are used to reduce the inter-cell interference.

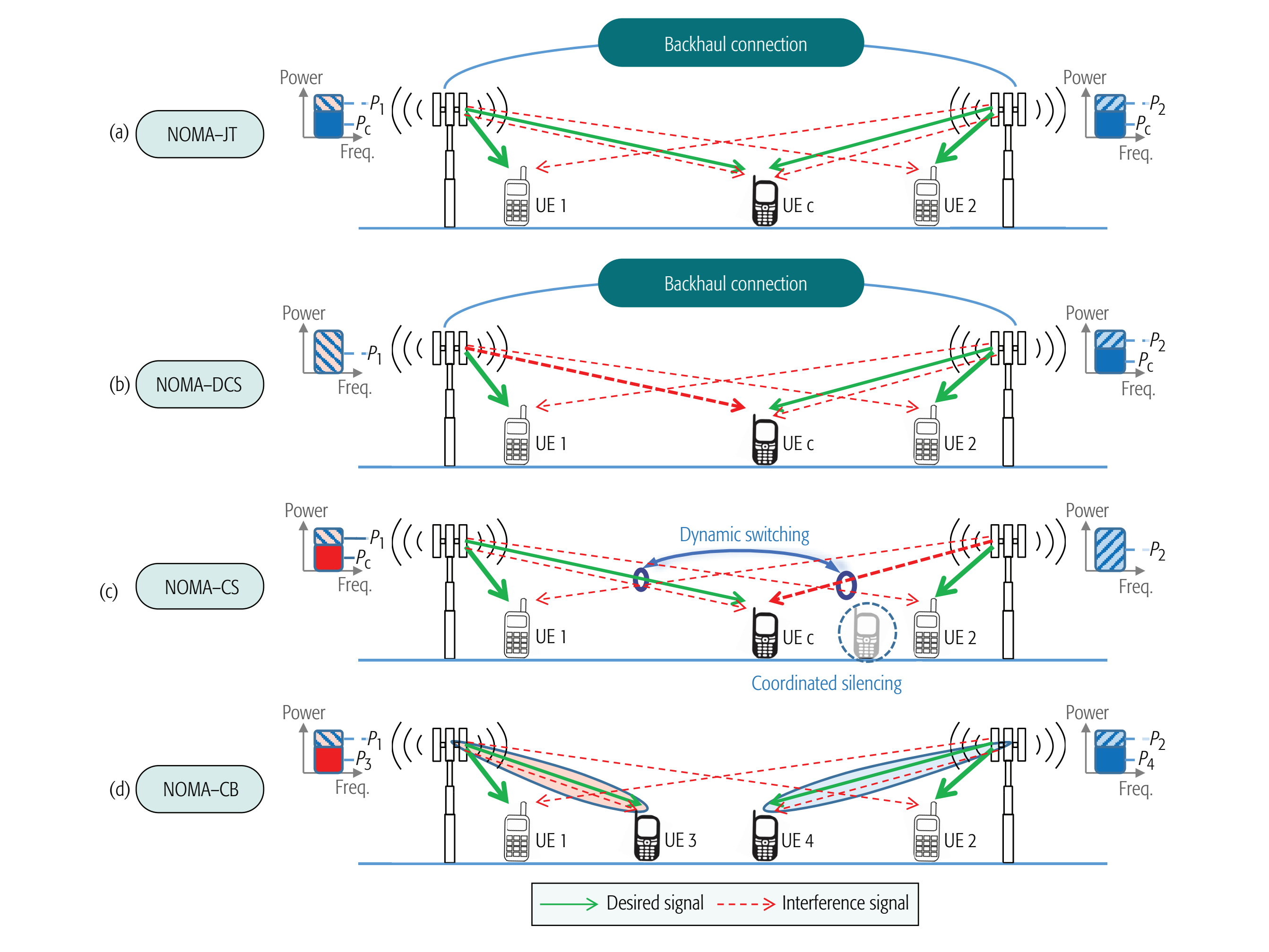


Figure 3. Multi-cell NOMA solutions: a) NOMA-JT; b) NOMA-DCS; c) NOMA-CS; d) NOMA-CB.

LTE-A systems under the name multi-user superposition transmission [9].

The performance of NOMA can be further boosted in multi-antenna networks. MIMO-NOMA solutions exploit *multiplexing* and *diversity* gains to improve outage probability and throughput, by converting the MIMO channel into multiple parallel channels [10].

Multi-Cell NOMA Solutions

In this section, we discuss recent research that combines interference management approaches with NOMA, called multi-cell NOMA. As illustrated in Fig. 2, ICI is the main issue in multi-cell NOMA networks, as it reduces a cell-edge user's performance. This is in contrast with single-cell NOMA, which aims at improving user-fairness. Multi-cell techniques are used to harness the effect of ICI.

Multi-cell techniques can be categorized into *coordinated scheduling/beamforming* (CS/CB) and *joint processing* (JP) [11]. This classification is based on whether the data messages desired by the users should be shared among multiple BSs or not. These techniques can be combined with NOMA. For NOMA-CS/CB, data for a user is only available at and transmitted from a single BS. In contrast, NOMA-JP relies on data sharing among more than one BS.

NOMA with Joint Processing

In NOMA-JP, the users' data symbols are available at more than one BS. Based on the number of active BSs that serve a user, we can further divide NOMA-JP into two classes: NOMA-joint transmission (JT) and NOMA-dynamic cell selection (DCS).

NOMA-JT: This approach requires multiple BSs to simultaneously serve a user using a shared wireless resource instead of acting as interference to each other. This significantly improves the quality of the received signal at cell-edge users at the cost of slightly diminished rates for cell-center users. This cooperative setting is similar to single-cell NOMA as the ICI for cell-edge users can be completely canceled using network MIMO techniques [12]. Such an approach usually relies on global channel state information (CSI) at all transmitters, which results in excessive backhaul overhead. To overcome the CSI sharing overhead for NOMA-JT, a coordinated superposition coding (CSC) scheme for a two-cell downlink network was introduced in [13]. In this scheme, each cell-center user is served by its corresponding BS while the cell-edge user is served by both BSs, as shown in Fig. 3. Specifically, two BSs transmit Alamouti coded signals to a cell-edge user to achieve a higher transmission rate, while each BS also transmits signals to the cell-center user. It

	NOMA-CS	NOMA-CB	NOMA-DCS	NOMA-JT
Number of transmission points	1	1	1 (dynamic)	≥ 2
Shared information	CSI, scheduling	CSI, BF	CSI, data	(CSI), data, BF
Backhaul type	Non-ideal	Non-ideal	Ideal	Ideal
Total number of supported users	$\ll 4K$	$4(K - 1)$	$3K$	$3K$ (or $4K$)
References		[14]	[13]	[12, 13]

Table 1. A comparison of different multi-cell NOMA solutions.

has been shown that the coordination between two cells allows NOMA to provide a common cell-edge user with a reasonable transmission rate without sacrificing cell-center users' rates. Let P and P_c be the powers of the cell-center and cell-edge users' messages per cell, respectively. Assume that $\gamma_{i,m} = |h_{i,m}|^2$ for $i \in \{1, 2, c\}$ and $m \in \{1, 2\}$, where $h_{j,m}$ and $h_{c,m}$ denote the channel coefficients to the cell-center user in cell j and the common cell-edge user from BS m, $\forall j, m \in \{1, 2\}$, respectively. The sum-rate of NOMA-JT given by $R_1 + R_2 + R_c$ (sum of rates for cell-center users R_1 and R_2, and a common cell-edge user R_c) where

$$R_1 = \mathcal{C}\left(\frac{\gamma_{1,1}P}{\gamma_{1,2}P+1}\right), R_2 = \mathcal{C}\left(\frac{\gamma_{2,2}P}{\gamma_{2,1}P+1}\right), \text{and}$$

$$R_\mathrm{c} = \min\left\{\begin{matrix} \mathcal{C}\left(\frac{(\gamma_{1,1}+\gamma_{1,2})P_c}{(\gamma_{1,1}+\gamma_{1,2})P+1}\right), \\ \mathcal{C}\left(\frac{(\gamma_{2,1}+\gamma_{2,2})P_c}{(\gamma_{2,1}+\gamma_{2,2})P+1}\right), \\ \mathcal{C}\left(\frac{(\gamma_{c,1}+\gamma_{c,2})P_c}{(\gamma_{c,1}+\gamma_{c,2})P+1}\right) \end{matrix}\right\}.$$

Note that the last term comes from the condition that a cell-edge user's message has to be decoded by that user and also cell-center users in both cells in order to operate SIC.

NOMA-DCS: In this case, the user's data is shared among multiple BSs, but it is transmitted only from one selected BS. Note that the transmitting BS can be dynamically changed over time by using order statistics. Suppose $|h_{c,2}|^2 > |h_{c,1}|^2$; then, BS 2 becomes the sole serving BS for a cell-edge user until the order statistics change. That is, only BS 2 employs a NOMA strategy to support a pair of cell-edge and cell-center users at the same time, while BS 1 serves only its corresponding cell-center user (Fig. 3). Since BS 1 employs OMA instead of NOMA, rate expressions for NOMA-JT, except for R_2, should be modified for NOMA-DCS as

$$R_1 = \mathcal{C}\left(\frac{\gamma_1 P}{\gamma_1(P+P_c)+1}\right) \text{and}$$

$$R_\mathrm{c} = \min\left\{\mathcal{C}\left(\frac{\gamma_2 P_c}{(\gamma_2+\overline{\gamma}_2)P+1}\right), \mathcal{C}\left(\frac{\gamma_2^c P_c}{(\gamma_1^c+\overline{\gamma}_2^c)P+1}\right)\right\}$$

since user 1 does not use SIC for NOMA transmission.

NOMA with Coordinated Scheduling/Beamforming

The designs of CS/CB for NOMA differ from those of JP in that the users' data are not shared among the BSs. However, the cooperating BSs still need to exchange global CSI and cooperative scheduling information via a standardized interface named X2. This may result in a non-negligible overhead, especially for high mobility cell-edge users. This subsection briefly discusses how to apply CS or CB to NOMA to tackle the ICI problem. An illustration of NOMA-CS/CB is shown in Fig. 3.

NOMA-CB: In this case, data for a user is only available at one serving BS, and the beamforming decision is made with coordination that relies on global CSI. The authors in [14] proposed two novel *interference alignment* (IA)-based CB methods in which two BSs jointly optimize their beamforming vectors in order to improve the data rates of cell-edge users by removing ICI. Both algorithms aim to choose the transmit/receive beamforming vectors to satisfy zero ICI as well as zero inter-cluster interference. These algorithms are termed interfering channel alignment (ICA)-based CB and IA-based CB. The former requires global CSI at the BS. However, the latter only requires the knowledge of cell-edge users' serving channels at the BS but with a slightly large number of antennas to compensate for the lack of interfering channels' knowledge. In particular, when the number of users is sufficiently large, it turns out that the number of extra antennas required for the latter scheme becomes negligible. Moreover, the transmit and receive beamforming vectors for uplink multi-cell NOMA also can be directly obtained by uplink-downlink duality.

NOMA-CS: The key idea of NOMA-CS is to allow geographically separated BSs to coordinate scheduling to serve NOMA users with less ICI so as to ensure the proper QoS of cell-edge users. To guarantee the required data rate of the cell-edge users in a cell, the adjacent BS may decide not to transmit a superimposed message to a set of NOMA users, but just a dedicated message to a single cell-center user as in Fig. 3. However, such a NOMA-CS scheme is formulated as a combinatorial optimization problem, which is NP-hard. Therefore, a simple scheduling algorithm is indispensable in order to determine a set of NOMA users scheduled in each BS within a certain scheduling interval.

A summary of different multi-cell NOMA techniques is provided in Table 1. In addition, we compare the number of users supported by different NOMA schemes according to the number of clusters in each cell and the number of BS/user antennas. For comparison, we consider two-cell scenarios and assume that each BS and user has K antennas. Each cell consists of K clusters each having two users. It should be highlighted that single-cell NOMA can support $2K$ users [10]. In contrast, single-cell OMA can serve only K users since the number of served users is limited by the number of antennas at the BS [4].

Practical Challenges for Multi-Cell NOMA

SIC Implementation Issues

As seen previously, SIC is at the heart of NOMA, and NOMA achieves the capacity region of the downlink and uplink channels (in a single-cell network) and the best known rate region in the multi-cell setting. However, SIC suffers from several practical issues, as described below.

Hardware Complexity: SIC implies that each user has to decode information intended for all other users before its own in the SIC decoding order. This causes the complexity of decoding to scale with the number of users in the cell. To reduce the complexity, we can divide users into multiple clusters and apply encoding/decoding within each cluster. Then, the complexity would be reasonable enough to be handled thanks to the advance of processor technologies in the past decades. In fact, 3GPP LTE-A/LTE-A Pro recently included a new category of relatively complex user terminal capabilities, named network assisted interference cancellation and suppression (NAICS).

Error Propagation: Error propagation means that if an error occurs in decoding a certain user's signal, all other users after this user in the SIC decoding order will be affected and their signals are likely to be decoded incorrectly. The problem can be compensated by using stronger codes provided that the number of users is not very large.[2]

Imperfect CSI

Without perfect CSI at the user side it is not possible to completely remove the effects of the other users' signals from the received signal, which results in error propagation. Moreover, without perfect CSI about the interfering links at the BS, a joint precoder that guarantees no ICI is not yet known. In this regard, new beamforming designs which are robust to CSI errors must be developed for multi-cell NOMA.

Multi-User Power Allocation and Clustering

Power allocation and clustering are important factors that determine the performance gain of NOMA. To explain the effect of these factors, consider the simple case of single-cell two-user NOMA and assume that βP and $(1 - \beta)P$ are the powers allocated to user 1 and user 2, respectively. As described earlier, by varying β, different points on the NOMA curve can be achieved. Therefore, β determines the rates for the users. This implies that with power allocation we can manage the system throughput and user-fairness. If there are more than two users in one cell, from the theory we know that all users' signals should be superimposed together; that is, having one cluster maximizes the system throughput . However, in practice, having only one cluster can result in serious performance degradation due to SIC error when there are many users in each cell.

A suboptimal, but more practical, solution is to have multiple clusters per cell. However, it is still very hard to find the optimal clustering for a given number of clusters and the optimal solution is unknown. In multi-cell networks, ICI comes in which makes clustering and power allocation even harder. It should be noted that in the multi-cell case, even when no SIC error is assumed, the optimal clustering and power allocation solutions are not known. Therefore, clustering and power allocation algorithms with reasonable complexity and good performance are inevitable to implement NOMA in practical cellular systems.

Operation with FFR

The basic idea of fractional frequency reuse (FFR) is to split a cell's bandwidth into multiple subbands and orthogonally allocate subbands for the cell-edge regions of the adjacent cells. This concept is in contrast with NOMA wherein orthogonalization is avoided due to its suboptimality. Despite being theoretically suboptimal, FFR is important as it offers a simple approach for ICI management without requiring CSI. Thus, it is important to investigate methods that can bring NOMA and FFR-based networks together. A simple idea to make use of both FFR and NOMA is to apply NOMA in the cell-center band and cell-edge band separately, which would pair cell-center users together (in the cell-center band) and cell-edge users together (in the cell-edge band). However, such users are not expected to have very different channel conditions, and any NOMA gain may not be noteworthy. Another idea is to pair a user from the cell-center region with a user from the cell-edge region in the cell-edge band to avoid ICI. However, such a pairing will reduce cell-edge users' rates as their specific bands can be shared by the cell-center users too, which sacrifices the cell-edge users' rates. This, in turn, deteriorates user-fairness.

Despite being theoretically suboptimal, FFR is important as it offers a simple approach for ICI management without requiring CSI. Thus, it is important to investigate methods that can bring NOMA and FFR-based networks together.

Security

The fact that in a NOMA-based transmission the user with better channel condition is able to decode the other user's signal brings new security concerns. Upper-layer security approaches (e.g., cryptography) are still relevant since only the legitimate user has a key to decode its message. Nonetheless, physical layer security schemes are of interest but cannot be easily applied to the new environment.

Performance of Multi-Cell NOMA

In order to observe the potential gain of NOMA, we perform a numerical analysis under a realistic multi-cell environment. In particular, we consider a two-cell downlink cellular network. As a performance metric, we use the cumulative distribution function (CDF) of the user throughput, and the individual user throughputs for the cell-center and cell-edge users. Detailed simulation parameters are provided in Table 2. In particular, the following schemes are considered: OMA, OMA-FFR, NOMA, NOMA-TDM, NOMA-JT, and NOMA-CB. In OMA-FFR, FFR is used in controlling ICI on top of OMA transmission. Due to the effect of FFR, the cell-edge users experience no ICI while the cell-center users receive interference from the other cell. In OMA, single-cell operation [10] is applied by treating all ICI as noise. Compared to OMA-FFR, ICI is a significant issue especially for cell-edge users, resulting in a severe SNR loss. Since it is inherently difficult to apply FFR in NOMA-based schemes as discussed previously, NOMA-TDM and NOMA schemes are considered. NOMA-TDM refers to a NOMA scheme that allows users in different cells to share one resource block via some form of orthogonalization, but NOMA simply acts as a single-cell operation [10] by treating the ICI as noise. In NOMA-CB and NOMA-JT, two BSs jointly optimize their beamforming vectors in order to mitigate ICI [13, 14].

[2] By making use of implementable near-capacity achieving AWGN channel codes (such as LDPC codes), we can get closer to the capacity region in practice.

> OMA and NOMA offer improved throughput for cell-center users due to the full usage of a resource block, but these techniques suffer from severe ICI, especially for the cell-edge users

Cell layout	2 Cells
Cell radius	0.25 Km
Path loss exponent	4
Channel model	Rayleigh fading model
Channel estimation	Ideal
Number of transmitter antennas	4
Number of receiver antennas	4
Number of clusters per cell	4
Number of users per cluster	2
Users' locations	Randomly generated and uniformly distributed within the cell
User pairing	Cell-center user from the disc with radius 0.125 Km; cell-edge user from the ring
Transmission power	10 W
Noise power spectral density	10^{-10} W/Hz
Maximum number of multiplexed UEs	1 (OMA), 2 (NOMA)

Table 2. Simulation parameters.

In Fig. 4a, we plot the performance of the cell-center and cell-edge users. Generally, the performance of OMA and NOMA decreases significantly with the location of cell-edge users since ICI mitigation is not considered. On the other hand, NOMA-TDM and OMA-FFR divide resources to support a multi-cell environment, and thus the rates of cell-edge users are improved compared to the single-cell operation schemes, such as OMA and NOMA. NOMA-CB can fully exploit all the resources to support all the users, and performs twice as well as NOMA-TDM. NOMA-JT performs similarly to NOMA-CB, but its gain increases as the cell-edge user gets closer to the border of the cell, because the cell-edge user can take advantage of the link from the neighboring BS to improve its SNR via data sharing. Note that the cell-edge user performance of OMA-FFR and OMA is even better than that of NOMA-CB when the location of the cell-edge user is relatively close to the BS, due to the inherently remaining inter-user interference of the cell-edge user from NOMA transmission [15]. This phenomenon is one motivation for NOMA to be implemented in practice.

In Fig. 4b, we plot the CDF of the user throughput. It can be seen that NOMA-CB and NOMA-JT achieve the best performance for any throughput because ICI is effectively controlled by exploiting the multi-cell NOMA transmissions. As a matter of fact, OMA and NOMA offer improved throughput for cell-center users due to the full usage of a resource block, but these techniques suffer from severe ICI, especially for the cell-edge users. In contrast, NOMA-TDM and OMA-FFR are deployed to overcome ICI by further splitting a resource block into two parts (i.e., two cells) while sacrificing cell-center users' throughput.

Conclusion

In this article, we have described the theory behind NOMA in single-cell (both uplink and downlink) and multi-cell networks. This has been followed by an up-to-date literature review of interference management techniques that apply NOMA in multi-cell networks. Numerical results have shown the significance of interference cancellation in NOMA. We have also highlighted major practical issues and challenges that arise in the implementation of multi-cell NOMA.

Acknowledgment

The work of W. Shin and J. Lee was supported in part by the Basic Science Research Program (NRF-2017R1A2B2007102) through the National Research Foundation of Korea funded by MSIP, INMAC, and BK21-plus. The work of M. Vaezi was supported by an NSERC PDF fellowship. The work of H. V. Poor was supported in part by the U.S. National Science Foundation under Grant EECS-1343210. D. J. Love was supported in part by the U.S. National Science Foundation under Grant CNS1642982.

References

[1] Y. Saito *et al.*, "Non-Orthogonal Multiple Access (NOMA) for Cellular Future Radio Access," *Proc. IEEE 77th Vehicular Technology Conf. (VTC Spring)*, 2013, pp. 1–5.

[2] L. Dai *et al.*, "Non-Orthogonal Multiple Access for 5G: Solutions, Challenges, Opportunities, and Future Research Trends," *IEEE Commun. Mag.*, vol. 53, no. 9, 2015, pp. 74–81.

[3] S. M. R. Islam *et al.*, "Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges," *IEEE Commun. Surveys Tutorials*, DOI:10.1109/COMST.2016.2621116.

[4] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge University Press, 2005.

[5] A. El Gamal and Y. H. Kim, *Network Information Theory*, Cambridge University Press, 2011.

[6] M. Vaezi and H. V. Poor, "Simplified Han-Kobayashi Region for One-Sided and Mixed Gaussian Interference Channels," *Proc. IEEE Int'l. Conf. Commun. (ICC)*, 2016, pp. 1–6.

[7] Y. Pang and M. Varanasi, "Approximate Capacity Region of the MAC-IC-MAC," arXiv preprint arXiv:1604.02234, 2016.

[8] Z. Ding, P. Fan, and H. V. Poor, "Impact of User Pairing on 5G Non-Orthogonal Multiple Access Downlink Transmissions," *IEEE Trans. Veh. Technol.*, vol. 19, no. 8, 2015, pp. 1462–65.

[9] 3GPP TD RP-150496, "Study on Downlink Multiuser Superposition Transmission," Mar. 2015.

[10] Z. Ding, F. Adachi, and H. V. Poor, "The Application of MIMO to Non-Orthogonal Multiple Access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, 2016, pp. 537–52.

[11] D. Lee *et al.*, "Coordinated Multipoint Transmission and Reception in LTE-Advanced: Deployment Scenarios and Operational Challenges," *IEEE Commun. Mag.*, vol. 50, no. 2, 2012, pp. 148–55.

[12] S. Han *et al.*, "Energy Efficiency and Spectrum Efficiency Co-Design: From NOMA to Network NOMA," *IEEE Multimedia Commun. Technical Committee E-Letter*, vol. 9, no. 5, 2014, pp. 21–24.

[13] J. Choi, "Non-Orthogonal Multiple Access in Downlink Coordinated Two-Point Systems," *IEEE Commun. Lett.*, vol. 18, no. 2, 2014, pp. 313–16.

[14] W. Shin *et al.*, "Coordinated Beamforming for Multi-Cell MIMO-NOMA," *IEEE Commun. Lett.*, vol. 21, no. 1, 2017, pp. 84–87.

[15] H. Tabassum *et al.*, "Non-Orthogonal Multiple Access (NOMA) in Cellular Uplink and Downlink: Challenges and Enabling Techniques," arXiv preprint arXiv:1608.05783, 2016.

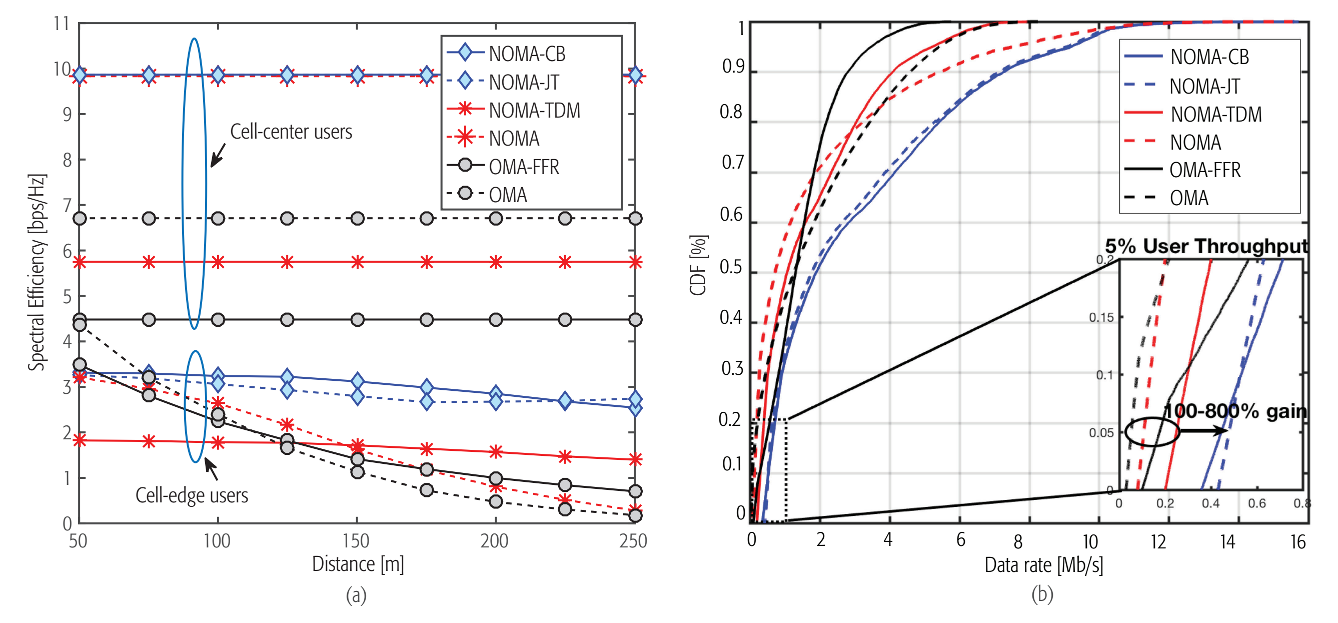


Figure 4. Performance comparison of different transmission schemes in multi-cell downlink networks: a) spectral efficiency as a function of cell-edge user's location; b) individual data rate CDF for random user deployments.

Biographies

Wonjae Shin [S'14] (wonjae.shin@snu.ac.kr) is a Ph.D. candidate in the Department of Electrical and Computer Engineering at Seoul National University, Korea. He was a visiting research scholar at Princeton University, Princeton, NJ, USA from 2016 to 2017. From 2007 to 2014, he was a member of technical staff at Samsung Advanced Institute of Technology (SAIT) and Samsung Electronics Co. Ltd. in Korea. He was awarded the SNU Best Ph.D. Dissertation Award in 2017, and the Gold Prize in the 2014 IEEE Student Paper Contest. He has been a program co-chair for the IEEE VTC 2017-Spring Workshop.

Mojtaba Vaezi [M'14] (mvaezi@princeton.edu) is an associate research scholar in the Department of Electrical Engineering at Princeton University. His research interests include the broad areas of wireless communications and information theory. He is an associate technical editor of *IEEE Communication Magazine* and an organizing member of NOMA workshops at VTC-Spring'17 and Globecom'17. He is a recipient of the McGill Engineering Doctoral Award, the IEEE Larry K. Wilson Student Activities Award in 2013, and the NSERC Postdoctoral Fellowship in 2014.

Byungju Lee [M'15] (byungjulee@purdue.edu) is a postdoctoral scholar in the School of Electrical and Computer Engineering at Purdue University, West Lafayette, IN, USA. From 2014 to 2015, he was a postdoctoral fellow in the Department of Electrical and Computer Engineering at Seoul National University, Seoul, Korea. He received B.S. and Ph.D. degrees from Korea University in 2008 and 2014, respectively. His current research focuses on physical layer system design for 5G wireless communications.

David J. Love [F'15] (djlove@purdue.edu) is a professor in the School of Electrical and Computer Engineering at Purdue University. His research interests include topics in signal processing and communications. His work has received numerous awards, including the 2016 IEEE Communications Society Stephen O. Rice Prize, the 2015 IEEE Signal Processing Society Best Paper Award, and the 2009 IEEE Transactions on Vehicular Technology Jack Neubauer Memorial Award.

Jungwoo Lee [SM'07] (junglee@snu.ac.kr) is a professor in the department of electrical and computer engineering at Seoul National University. His research interests include wireless communications, information theory, distributed storage, and machine learning. He has been an editor for *IEEE Wireless Communications Letters* since 2017, and was an associate editor for *IEEE Transactions on Vehicular Technology* from 2008 to 2011. He received the Qualcomm Dr. Irwin Jacobs award in 2014 for his contributions in wireless communications.

H. Vincent Poor [F'87] (poor@princeton.edu) is the Michael Henry Strater University Professor of Electrical Engineering at Princeton University. His interests include information theory and signal processing, with applications in wireless networks and related fields. He is an IEEE Fellow, a Member of the National Academy of Engineering and the National Academy of Sciences, and a Foreign Member of the Royal Society. He received the Marconi and Armstrong Awards of the IEEE Communications Society in 2007 and 2009, respectively. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal and honorary doctorates from several universities.

Wireless-Optical Network Convergence: Enabling the 5G Architecture to Support Operational and End-User Services

Anna Tzanakaki, Markos Anastasopoulos, Ignacio Berberana, Dimitris Syrivelis, Paris Flegkas, Thanasis Korakis, Daniel Camps Mur, Ilker Demirkol, Jesús Gutiérrez, Eckhard Grass, Qing Wei, Emmanouil Pateromichelakis, Nikola Vucic, Albrecht Fehske, Michael Grieger, Michael Eiselt, Jens Bartelt, Gerhard Fettweis, George Lyberopoulos, Eleni Theodoropoulou, and Dimitra Simeonidou

The authors present a converged 5G network infrastructure and an overarching architecture to jointly support operational network and end-user services, proposed by the EU 5G PPP project 5G-XHaul. The 5G-XHaul infrastructure adopts a common fronthaul/backhaul network solution, deploying a wealth of wireless technologies and a hybrid active/passive optical transport, supporting flexible fronthaul split options.

Digital Object Identifier:
10.1109/MCOM.2017.1600643

Abstract

This article presents a converged 5G network infrastructure and an overarching architecture to jointly support operational network and end-user services, proposed by the EU 5G PPP project 5G-XHaul. The 5G-XHaul infrastructure adopts a common fronthaul/backhaul network solution, deploying a wealth of wireless technologies and a hybrid active/passive optical transport, supporting flexible fronthaul split options. This infrastructure is evaluated through a novel modeling. Numerical results indicate significant energy savings at the expense of increased end-user service delay.

Introduction

The enormous growth of mobile data predicted is attributed to the rapidly increasing:

- Number of network-connected end devices
- Internet users with heavy usage patterns
- Broadband access speed
- Popularity of applications including cloud computing, video, gaming, and so on

Traditional radio access networks (RANs), where baseband units (BBUs) and radio units (RUs) are co-located, cannot meet this massive foreseen growth. This is attributed to high capital and operational costs associated with the lack of resource sharing and modularity, reduced agility and scalability, as well as inefficient energy management.

Cloud RANs (C-RANs) propose to overcome these limitations, by supporting connection of access points (APs), known as RUs, to a BBU pool hosted in a central unit (CU) through a set of transport links. These links are referred to as fronthaul (FH). Currently, interfacing between RUs and a CU is enabled through the adoption of standards such as the common public radio interface (CPRI). The RU wireless signals are commonly transported over an optical FH network, using either digital transmission (e.g., CPRI) or analog transmission (radio-over-fiber). The adoption of CPRI-type solutions enables consolidation of a larger number of BBUs per CU by extending the transport network range. However, C-RAN requires very high transport bandwidth due to the traffic volume created by the sampled radio signals transported to the CU and the very tight delay and synchronization specifications [1]. Existing millimeter-wave (mmWave) E-Band and optical transport solutions supporting traditional backhaul (BH) requirements are based on different flavors of passive optical networks (PONs) and 10GE technologies. Considering that in fifth generation (5G) environments these transport solutions will also need to offer FH capabilities, it is clear that they will be unable to offer the required capacity for both BH and FH services. To take advantage of the benefits and address the challenges associated with C-RAN, equipment vendors are expanding their FH solutions adopting advanced wireless technologies; for example, sub-6GHz and 60 GHz bands, including advanced beam tracking and multiple-input multiple-output (MIMO) techniques; and new flexible and dynamic wavelength-division multiplexing (WDM) optical networks [2]. These are also enhanced with novel control and management approaches to enable increased granularity, end-to-end optimization, and guaranteed quality of service (QoS).

To facilitate CRAN's technical feasibility and benefit from its coordination and pooling gains, there is a need to relax the FH requirements. In view of this, solutions proposing FH compression and alternative architectures relying on flexible functional splits (Fig. 2) have been reported [3, 4]. The concept of flexible splits relies on transferring some of the processing functions away from the RU and locating these centrally at a CU. These functions are commonly performed through dedicated and specific-purpose hardware, with significant installation, operational, and administrative costs. To address these issues, the concept of network softwarization enabling migration from traditional closed networking models to an open reference platform able to instantiate a variety of network functions has been proposed recently.

Anna Tzanakaki is with the University of Bristol and National and Kapodistrian University of Athens; Markos Anastasopoulos and Dimitra Simeonidou are with the University of Bristol; Ignacio Berberana is with Telefonica Investigacion Y Desarrollo; Dimitris Syrivelis, Paris Flegkas, and Thanasis Korakis are with the University of Thessaly; Daniel Camps Mur is with i2CAT Foundation; Ilker Demirkol is with Univeristat Politecnica de Catalunya; Jesús Gutiérrez and Eckhard Grass are with IHP GmbH; Qing Wei, Emmanouil Pateromichelakis, and Nikola Vucic are with Huawei Technologies Duesseldorf; Albrecht Fehske and Michael Grieger are with Airrays GmbH; Michael Eiselt is with ADVA Optical Networking SE; Jens Bartelt and Gerhard Fettweis are with Technische Universität Dresden; George Lyberopoulos and Eleni Theodoropoulou and COSMOTE Mobile Communications S.A.

0163-6804/17/$25.00 © 2017 IEEE

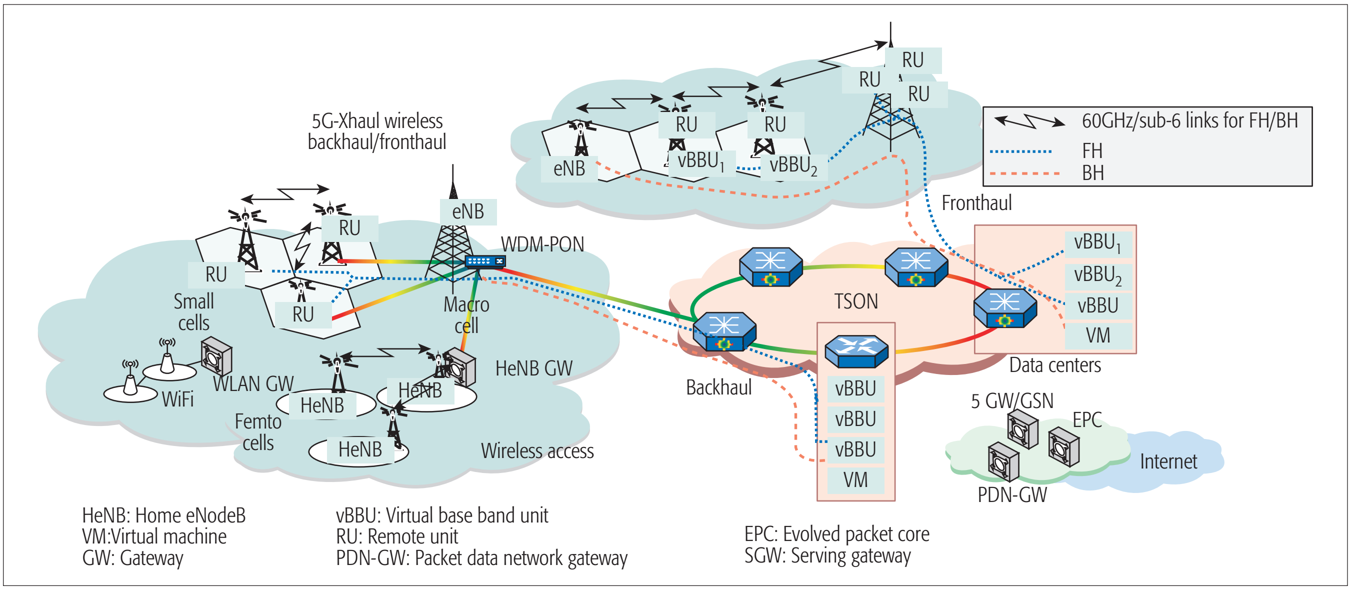


Figure 1. The 5G-XHaul Physical Infrastructure: FH and BH services are provided over a common wired/wireless network infrastructure. In the FH case, parts of the BBU processing can be performed locally and some parts remotely at the DCs enabling the C-RAN flexible split paradigm. BBUs are executed in general purpose servers in the form of virtual entities. BH services interconnect end-users with Virtual Machines hosted in the DCs.

A typical example includes the OpenAirInterface (OAI), an open source 4G/5G radio stack able to be executed on general-purpose servers hosted in data centers (DCs) [5]. Such open source frameworks are still in early development stages and do not allow execution of more complex functionalities such as flexible RAN splits. In this study, the concept of flexible functional splits is addressed by appropriately combining servers with low processing power (cloudlets) and relatively large-scale DCs placed in the access and metro domains, respectively. The remote processing requirements associated with some of the functional split options impose the need for a high bandwidth transport interconnecting RUs and the CU. On the other hand, the variability of remote processing requirements across the various split options introduce the need for a transport network that offers finely granular and elastic resource allocation capabilities.

Addressing these challenges, we propose a network solution that converges heterogeneous network domains deploying optical and wireless technologies together with compute resources in a common 5G infrastructure. This infrastructure, developed in the framework of the EU 5G Public Private Partnership (5GPPP) project 5G-XHaul, will support both the operational network as well as fixed and mobile end-user services. Operational network services refer to services required for the operation of the 5G infrastructure (e.g., FH services offered to infrastructure operators/providers). On the other hand, end-user services refer to services provided to end users (content delivery, gaming, etc.) that in 5G environments require BH connectivity, referred to as BH services. The main technical innovations of the proposed solution include:

- An architectural framework aligned with the software defined networking (SDN) open reference architecture [6] and the European Telecommunications Standards Institute (ETSI) network functions virtualization (NFV) standard to jointly support FH and BH services as well as the adoption of flexible functional split options

This is a key innovation of the proposed architecture compared to contemporary LTE-Advanced (LTE-A) systems where FH and BH services are supported by separate and dedicated networks, while network control and management is closed.

- A novel data plane design that converges heterogeneous wireless and optical solutions
- A novel modeling framework adopting multi-objective optimization (MOP) techniques to evaluate the proposed architectural approach

This modeling framework focuses on optimal FH and BH service provisioning, with the overall objective to maximize the infrastructure energy efficiency and minimize end-to-end service delays.

Overview of the 5G-XHaul Architecture

Data Plane Architecture

The 5G-XHaul data plane design considers converged optical and wireless network domains in a common 5G infrastructure supporting both transport and access. In the wireless domain, a dense layer of small cells can be wirelessly backhauled through mmWave and sub-6 GHz technologies. Alternatively, small cells can be connected to a CU through the 5G-XHaul optical network. This adopts a hybrid approach combining a dynamic and elastic frame-based optical network solution with enhanced capacity WDM PONs [7] to support the increased transport requirements of 5G environments in terms of granularity, capacity, and flexibility.

Given that 5G-XHaul proposes the adoption of C-RAN to overcome traditional RAN limitations (Fig. 1), it introduces the need to support new operational network services (FH) over the transport network. These emerge from the need to connect densely distributed RUs with the CU, meeting very tight latency and synchronization

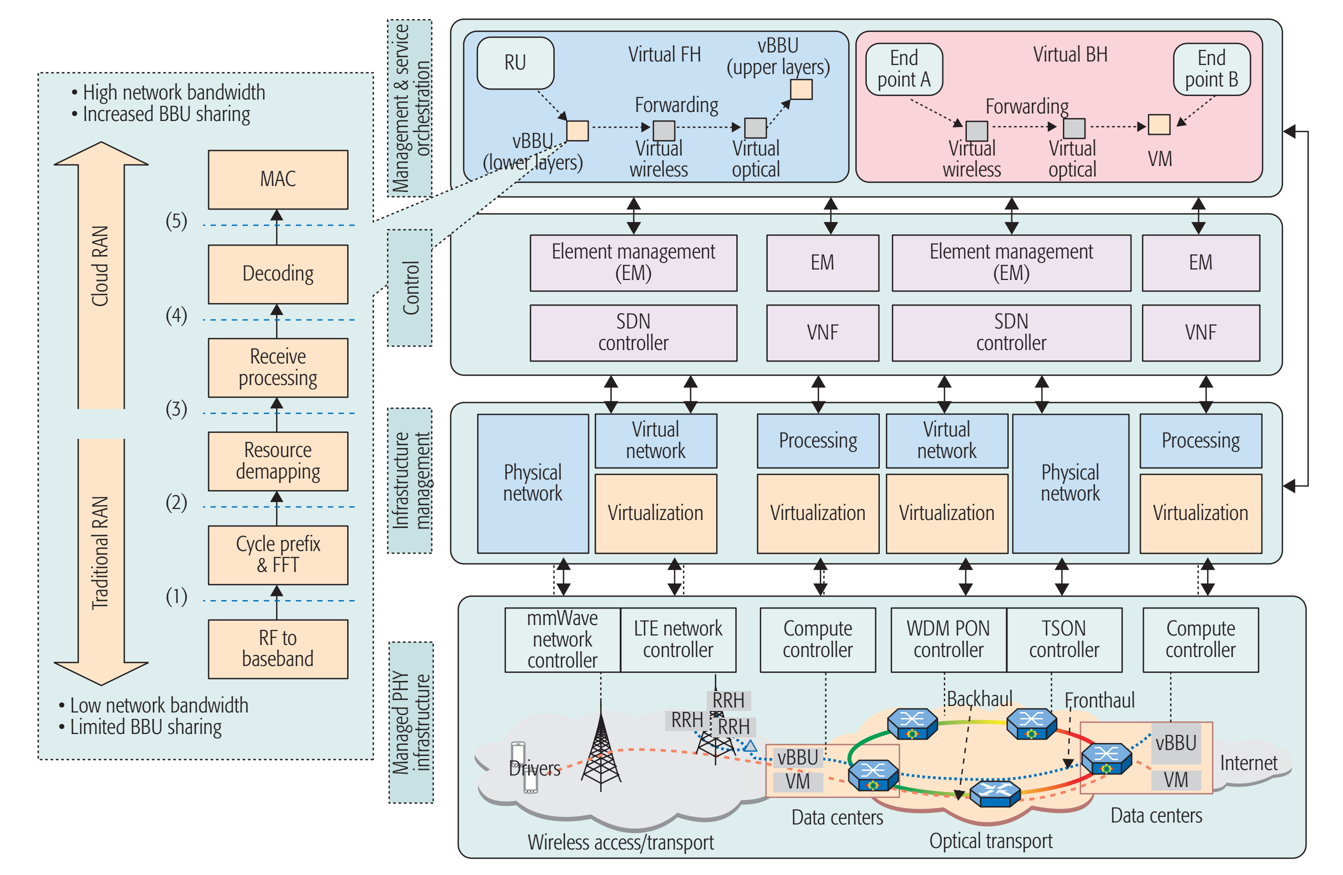


Figure 2. The overall overarching architecture supporting functional split processing [3, 4].

requirements. To maximize coordination and resource sharing gains, 5G-XHaul proposes to support BH and FH jointly in a common infrastructure, thus achieving improved efficiency and management simplification leading to measurable benefits in terms of cost, scalability, and sustainability. Aiming to address the C-RAN challenges described above, 5G-XHaul proposes the adoption of flexible split options that can relax the tight transport requirements in terms of capacity, delay, and synchronization. Figure 2 shows the range of optimal split options that span between the traditional distributed RAN case, where all processing is performed locally at the AP, to the fully centralized C-RAN case, where all processing is allocated at the CU. All other options allow allocating some processing functions at the RU, while the remaining processing functions are performed remotely at the CU. The optimal allocation of processing functions to be executed locally or remotely (i.e., the optimal split) can be decided based on factors such as transport network characteristics, network topology, and scale as well as type and volume of services.

A key enabler of the 5G-XHaul data plane (Fig. 1) is the hybrid (passive-active) optical network transport that jointly supports FH and BH services offering the required connectivity, capacity, and flexibility. The passive solution employs WDM-PONs, while the active solution adopts the highly versatile time-shared optical network (TSON) [7] extended to support novel features offering fine bandwidth granularity (variable length optical frames) and elastic bandwidth allocation capabilities.

Given the technology heterogeneity supported by the 5G-XHaul data plane, a critical function of the converged infrastructure is interfacing between technology domains. Interfaces are responsible for handling protocol adaptation as well as mapping and aggregation/de-aggregation of traffic across domains. Different domains (wireless/optical) may adopt different protocol implementations and provide very diverse levels of capacity (megabits per second for the wireless domain to tens of gigabits per second for TSON), granularity (kilobits per second for the wireless domain to 100 Mb/s for TSON), and so on. A key challenge also addressed by these interfaces is mapping of different QoS classes across different domains as well as flexible scheduling enabling QoS differentiation mechanisms. More specifically, at the optical network ingress point (e.g., TSON edge node), the interfaces receive traffic frames generated by fixed and mobile users and arrange them to different buffers. The incoming traffic is aggregated into optical frames and is assigned to suitable time slots and wavelengths according to the adopted queuing policy before transmission in the TSON domain. For FH traffic a modified version of the CPRI protocol supporting the concept of functional splits (eCPRI) has been adopted. Note that due to the large variety of technologies involved in 5G, these interfaces need to support a wide range of protocols and technology solutions and execute traffic forwarding decisions at wire-speed. This requires the development of programmable network interfaces combining hardware-level performance with software flexibility. The reverse function is performed at the egress TSON edge node. More information regarding interfacing of wireless and optical domains is available in [7, 13].

Overarching Layered Architecture

Managing and operating heterogeneous infrastructures integrating a variety of optical and wireless technologies and domains such as the 5G-XHaul infrastructure present several challenges. To address these we propose the adoption of the integrated SDN/NFV paradigm. This will take advantage of the separation of the control and data planes offered by SDN and the depoyment of the variety of NFV elements. Through this integration the benefits of the control and the holistic network view of SDN will be combined with the flexibility to provision services by composing service chains (SCs) through orchestrated network functions. SDN/NFV integration allows SDN controllers to control the virtual network functions (VNFs) [11] enabling on-demand resource allocation for dynamically changing workloads [6]. SDN network elements may correspond to both physical network functions (PNFs) and VNFs if they are implemented in virtualized environments, as software running on general-purpose hardware platforms [6]. The virtualization of network elements enables flexible allocation of data plane resources according to network applications' requirements. On the other hand, SCs offering orchestrated service provisioning over heterogeneous environments are considered to be a possible network application, which can include SDN controller functions or interact with SDN controllers to provide VNFs.

The details of the 5G-XHaul overarching architecture adopting the integrated SDN/NFV paradigm to facilitate management and operation of the heterogeneous physical infrastructure (PI) are illustrated in Fig. 2.

The infrastructure management layer (IML) manages the different technology domains. This layer is responsible to enable multi-tenant operation through cross-domain slicing and virtualization facilitating joint FH and BH services over the common infrastructure. Information retrieval and communication between domains is handled by network and compute controllers located at this layer, enabling resource abstraction and virtualization. Therefore, IML supports traditional management of the PI together with advanced features required for virtualization and virtual resource management functions.

The control layer (CL) is responsible for cross-domain orchestration of virtual infrastructures (VIs) and PIs, created and exposed by the IML having an overall view of all network domains. The CL provides end-to-end connectivity services in the form of SCs deploying converged control and management procedures with guaranteed QoS. The CL supports configuration of both virtualized and non-virtualized heterogeneous resources as well as legacy devices through a set of distributed SDN controllers, and facilitates the development of enhanced VNFs to operate the 5G infrastructure seamlessly.

The management and service orchestration layer (MSOL) handles orchestration requirements for the delivery of network and compute services as well as composition and provisioning of SCs in multi-tenant environments deploying VNFs. The MSOL is also responsible for supporting interoperability with legacy software and hardware.

Use Case: Joint Optimization of FH/BH

As already discussed, the 5G-XHaul data plane architecture can jointly support FH and BH services, adopting a hybrid optical transport, integrating passive and active optical networks. The higher layers of the architecture that facilitate access and management of both network and compute resources also play a key role. The ability of the IML to create VI slices across heterogeneous domains and to expose these to the upper layers is an instrumental architectural tool, facilitating the delivery of FH and BH services. Identifying optimal VIs in terms of both topology and resources includes:

- Ordering, referred to as SC, of the relevant functions (VNF or PNF) that need to be applied to the traffic flows traversing the VIs
- Estimating the virtual resources required to support SC and executing the corresponding applications over the PI
- Mapping of the virtual resources to the physical resources

This process is shown in the upper part of Fig. 2. In this example two VIs, corresponding to different tenants, are able to independently support FH and BH functions over a common infrastructure.

In 5G-XHaul we assume a multi-technology transport network interconnecting RUs and end users with a set of general-purpose servers hosted by the CU (as showcased, e.g. by Alcatel-Lucent, Intel, China Mobile, and Telefónica at Mobile World Congress 2015). To support virtual FH (VFH) services, over the 5G-XHaul infrastructure, RU demands are forwarded to a shared CU, hosting a set of sliceable and virtualized servers for processing. Compute resources are responsible for executing the various BB functions in a predefined order (Fig. 2, left). Based on the split option adopted, these functions can be partly executed locally at the RU or centrally at the CU. The split choice dictates the processing allocation to local and central compute resources, and enforces the corresponding SC graph. A graphical representation of a typical virtual BH (VBH) service that supports content delivery (a content delivery network, CDN) to end users is shown in Fig. 2. The SC graph indicates that the VBH service allows mobile traffic, generated at the wireless access domain, to traverse a hybrid multihop wireless/optical transport network before it reaches the compute resources.

To evaluate the performance of this type of infrastructure and the proposed architecture, we have developed a mathematical framework, based on MOP, for the integrated wireless and optical network domains, considering the data plane described earlier and the details of the compute resources required. Our study focuses on optimal planning of VFH and VBH infrastructures in terms of both topology and resources, considering overall power consumption and end-to-end delays. To identify the best performing FH and BH VIs, detailed power consumption and end-to-end delay models are considered. These models

Managing and operating heterogeneous infrastructures integrating a variety of optical and wireless technologies and domains such as the 5G-XHaul infrastructure presents several challenges. To address these we propose the adoption of the integrated SDN/NFV paradigm.

As the optical transport requires a very small amount of power to operate, higher pooling gains are expected when C-RAN solutions are adopted compared to traditional RANs. The impact of centralization is expressed through transport network overloading introduced by FH services, leaving limited resources for BH services.

describe the details of the optical and wireless network as well as the compute domains and the associated interfaces, [9]. The results obtained focus on the specific use case of joint FH and BH optimization, assuming delivery of CDN services to the end users.

The joint VFH/VBH design problem also considers a set of constraints ensuring efficient and stable operation of the planned Vis, summarized below.

•VFH and VBH infrastructures have specific requirements. In response to this, different VNFs are grouped and orchestrated in the form of SCs with specific processing and network requirements. To realize an SC, sufficient network and processing capacity must be allocated to the planned VIs for the interconnection and deployment of VNFs. The order of VNF processing is defined by the corresponding SC.

•Reservation of physical resources to support SC depends on the users' mobility model assumed, the size of the wireless cells, and the traffic model adopted. Ideally, 100 percent overprovisioning of both network and compute resources across neighboring cells can guarantee seamless handovers. However, to improve resource efficiency, the reservation of resources residing in adjacent cells can be linked to handoff probability. The amount of resources leased in the wireless domain is assumed to be an increasing function of the handoff probability [12]. Given that both the RUs and the end users need to be supported by remotely located compute resources, the additional resource requirements also propagate in the transport network and the compute domain.

•The VI planning process considers a number of functions across different domains including flow conservation, mapping, aggregation, and deaggregation of traffic.

•Given that both FH and BH services require compute processing, the associated impact on the overall infrastructure evaluation is considered. Therefore, the traffic associated with these services is mapped, not only to network resources, but also to compute resource requirements. This introduces an additional constraint, linked with the conversion of network-to-compute resource requests. To achieve this, a mapping parameter, defined as the "network-to-compute" parameter, is introduced providing the ratio of network requirements (in megabits per second) and computational requirements (in operations per second [OPS]), of a specific service demand. This parameter takes high values for cloud services requiring high network bandwidth and low processing capacity (e.g., video streaming). On the other hand, it takes low values for tasks requiring intensive processing and low network bandwidth (e.g., data mining, the Internet of Things). Regarding BH services, the Standard Performance Evaluation Corporation recently established the Cloud subcommittee to develop benchmarks able to measure these parameters. Taking a similar approach, the authors in [14] measured the average ratio between computational and network bandwidth requirements for various big data analytics workloads. Similarly, FH services require specific computing resources to support BB processing. The processing power depends on the details of the BBU [3, 4] including processing tasks related to fast Fourier transform (FFT), error correction, processing-resource mapping/de-mapping, and so on. calculated in Giga OPS (GOPS). The resulting processing power depends on the LTE system configuration [4].

•Our analysis takes into consideration end-to-end delays for specific services (e.g., FH or real-time BH services). In highly loaded heterogeneous networks, such as the 5G-XHaul solution, end-to-end delay can be greatly influenced by queuing delays associated with the interfaces across the infrastructure domains. In this context, the choice of suitable queuing and scheduling policies at the interfaces offers significant delay benefits. Traditionally, these systems can be mathematically modeled applying queuing theory and open/closed mixed queuing networks. However, such a model is not able to adhere to the strict FH latency constraints. To address this issue, the queuing delays for the VFH are modeled under worst case operational conditions using network calculus theory [13].

Flexible processing splits refer to the choice of a single split option for every time instance. Once the split option has been selected, the corresponding SC is applied across the network, deploying specific network and compute resources as dictated by the relevant split [3, 4].

Depending on the functional split, some of the processing is performed by compute resources either at a local cloudlet [12] c, $c \in C$ (C denotes the set of cloudlets) with cost w_c per GOPS or at a remote regional DC s, $s \in S$ (S denotes the set of DCs) with cost w_s per GOPS. Assuming that the cost for operating FH capacity $u_{FH,e}$ of physical link e $\in$ ε (ε: set of physical links) is w_e, and $\pi_{FH,s}$ $\pi_{FH,c}$ are the BBU processing capacities at the remote server and the cloudlet, respectively, the optimal VFH infrastructure is determined by minimizing the following cost:

$$\min \mathcal{VFH}(u,\pi) = \sum_{e\in\varepsilon} w_e u_{FH,e} + \sum_{s\in S} w_s \pi_{FH,s} + \sum_{c\in C} w_c \pi_{FH,c} \tag{1}$$

subject to the constraints described above.

As the optical transport requires a very small amount of power to operate, higher pooling gains are expected when C-RAN solutions are adopted compared to traditional RANs. The impact of centralization is expressed through transport network overloading introduced by FH services, leaving limited resources for BH services. In view of this we set a secondary optimization objective with the aim to minimize BH end-to-end delay, subject to demand processing and capacity constraints:

$$\min \mathcal{VBH}(u,\pi) = \sum_{e\in\varepsilon} \frac{1}{\mathcal{U}_c - u_{FH,e} - u_{BH,e}} + \sum_{s\in S} \frac{1}{\Pi_s - \pi_{FH,s} - \pi_{BH,s}} + \sum_{c\in C} \frac{1}{\Pi_c - \pi_{FH,c} - \pi_{BH,c}} \tag{2}$$

where $u_{BH,e}$, $\pi_{BH,s}$ represent the BH related network and server capacity, respectively, $\mathcal{U}_e$ is the

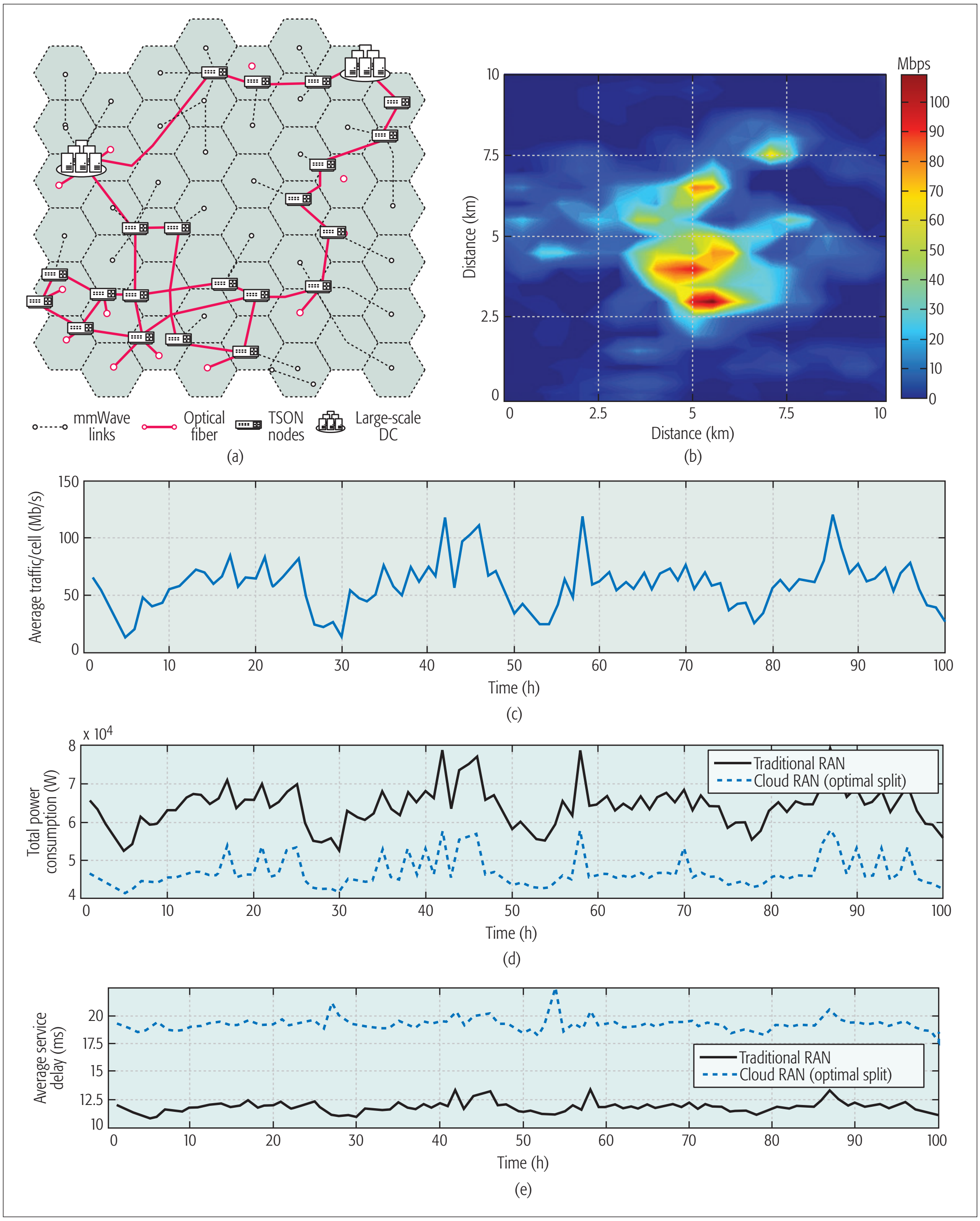


Figure 3. a) Bristol 5G city network topology with mmWave backhauling; b) snapshot of spatial traffic load; c) average traffic/BS based on the dataset [10] during 8/2012; d)–e) total power consumption and total service delay over time for the traditional RAN.

Given that both FH and BH services require compute processing, the associated impact on the overall infrastructure evaluation is considered. Therefore, the traffic associated with these services is mapped, not only to network resources, but also to compute resource requirements. This introduces an additional constraint, linked with the conversion of network-to-compute resource requests.

total capacity of e, and $\prod_s$, $\prod_c$ are the total processing capacity of the DC s and the cloudlet c, respectively.

The MOP described through Eqs. 1 and 2 can be written as min $\mathcal{F}(u, \pi) = [\mathcal{VFH}(u,\pi), \mathcal{VBH}(u, \pi))]$ subject to the previously discussed constraints. This problem is then transformed from an MOP into a single objective problem using the Pascoletti-Serafini scalarization technique [10] and solved using Lagrangian relaxation. In the following section, the performance of the overall architecture is evaluated in terms of power consumption and service delay adopting a realistic network topology and actual traffic statistics [15].

Performance Evaluation

The network topology assumed is the Bristol 5G city infrastructure (Fig. 3a). In this infrastructure a set of 50 APs are evenly distributed across a 10 × 10 km^2 area. APs are backhauled through microwave point-to-point links, and TSON is adopted for the optical transport. TSON deploys a single fiber per link, 4 wavelengths of 10 Gb/s each per fiber, and minimum bandwidth granularity of 100 Mb/s. In the present study, w_e is associated with the power consumption of link e ∈ ε. Power consumption figures for TSON can be found in [7, 13]. The microwave transceivers considered have 2 Gb/s bandwidth, and their power consumption is 45 W (Huawei OptiXRTN310).

A novel MOP modeling framework has been developed to evaluate the performance of the 5G-XHaul architecture taking into consideration the joint support of FH and BH services. Our modeling results show that the proposed architecture can offer significant benefits in terms of energy consumption but at the expense of end-user service delays.

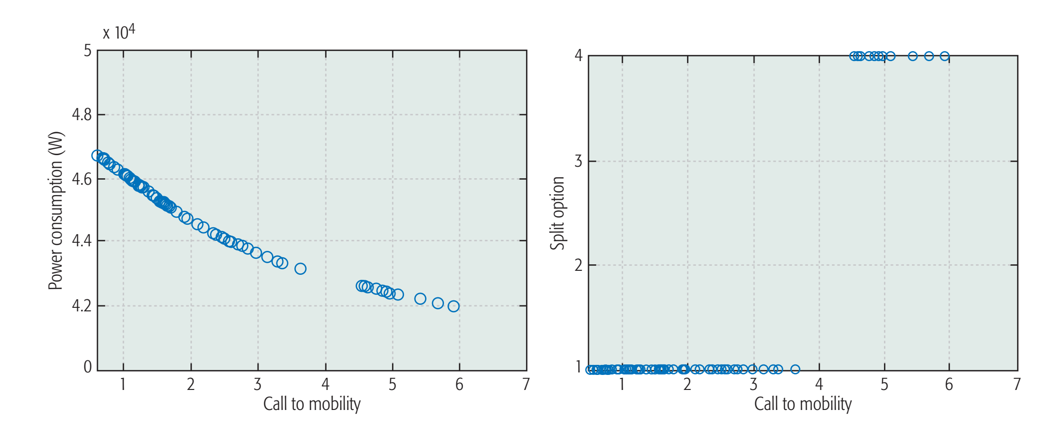


Figure 4. Impact of mobility on a) power consumption; b) optimal split option (load 18 Mb/s/cell).

Furthermore, a 2 × 2 MIMO scheme with adaptive number of transmission elements, carrier frequency 2.6 GHz, 20 MHz bandwidth adjustment, and capacity per cell up to 201.6 Mb/s has been considered. Mobile users are distributed and generate traffic over the serviced area according to real datasets reported in [15] (Fig. 3b). The following two scenarios are studied:

- "*Traditional RAN*," where power consumption per AP ranges between 600 and 1200 W under idle and full load conditions, respectively, and commodity servers are used to support CDN service
- "*C-RAN with virtual BBUs (vBBUs)*," where commodity servers are used to support both FH (through the creation of vBBUs [1]) and BH CDN services

For *C-RAN with vBBUs*, where optimal split options are deployed, two types of servers have been considered: a) small-scale commodity servers (cloudlets) close to the APs and b) commodity servers hosted by large-scale DCs (Fig. 3a) with an average cost equal to 2 W/GOPS and 1.6 W/GOPS, respectively. Details regarding the numerical values used in the simulations are provided in [7]. Although both types of servers can provide the necessary processing power for C-RAN and CDN services, large-scale DCs provide superior performance per Watt compared to cloudlets. Figure 3d shows that significant energy savings (ranging between 60–75 percent) can be achieved adopting the C-RAN approach using the integrated wireless-optical infrastructure, compared to traditional RAN. However, due to sharing of network resources between BH services and high-priority FH services, C-RAN leads in increased BH service delays that remain below 25 ms (Fig. 3e). On the other hand, traditional RAN provides minimum end-to-end BH service delays, as no sharing with FH services is required, but at the expense of increased power consumption due to the limited BBU sharing.

The impact of mobility on the total power consumption and the optimal split option adopted is shown in Figs. 4a and 4b, respectively. The call-to-mobility factor is defined as the ratio of the service holding time over the cell residence time [13], with low call-to-mobility factor values indicating high degree of mobility. It is known that high degree of mobility introduces additional resource requirements in the wireless domain. To ensure seamless end-to-end connectivity between end users, RUs, and compute resources, these additional resource requirements also propagate across the transport network and the compute domains. In Fig. 4b it is observed that lower split options are beneficial for higher mobility, enabling a larger number of BB processing tasks to be offloaded to remote DCs. Given that BB processing requirements increase with mobility, a higher degree of centralization benefits the system, due to the increased consolidation and improved performance per Watt that large-scale remote DCs offer, compared to local cloudlets.

Figure 5a illustrates the impact of service requirements in terms of network and compute resources on the optimal split option adopted. CDN services with high network-to-compute ratios (e.g., video analytics) require significant network resources to operate, leading to overutilization of transport capacity. This effect is counter-balanced by the selection of higher split options (i.e., options 3, 4) that require lower bandwidth for the interconnection of RUs with CUs compared to the bandwidth requirements of lower split options (i.e., options 1, 2). The impact of the traffic load on the total power consumption is illustrated in Fig. 5b. As expected, for higher traffic load, the total power consumption increases, and a step-like increase is observed, above 45 Mb/s per cell traffic load. Beyond this threshold, the preferable system split option becomes split 4 (rather than split 1), and a large number of cloudlets per geographic region are activated to support BB processing requirements.

Finally, the impact of the relative processing and transport network cost, on the optimal split option, is illustrated in Fig. 6. The relative local to remote processing cost, is defined as the ratio of the power consumed for data processing at the local cloudlet over the power consumed for processing of the same data remotely at large scale DCs. It is seen that increasing this ratio makes it beneficial to perform more processing functions at large-scale remote DCs. Thus, a lower split option is preferable. To include the impact of the network cost in this analysis, the end-to-end transmission cost is also plotted in Fig. 6. As the transmission cost increases (higher number of wireless hops), it is beneficial to perform more processing functions at the local cloudlets and hence adopt a higher split option.

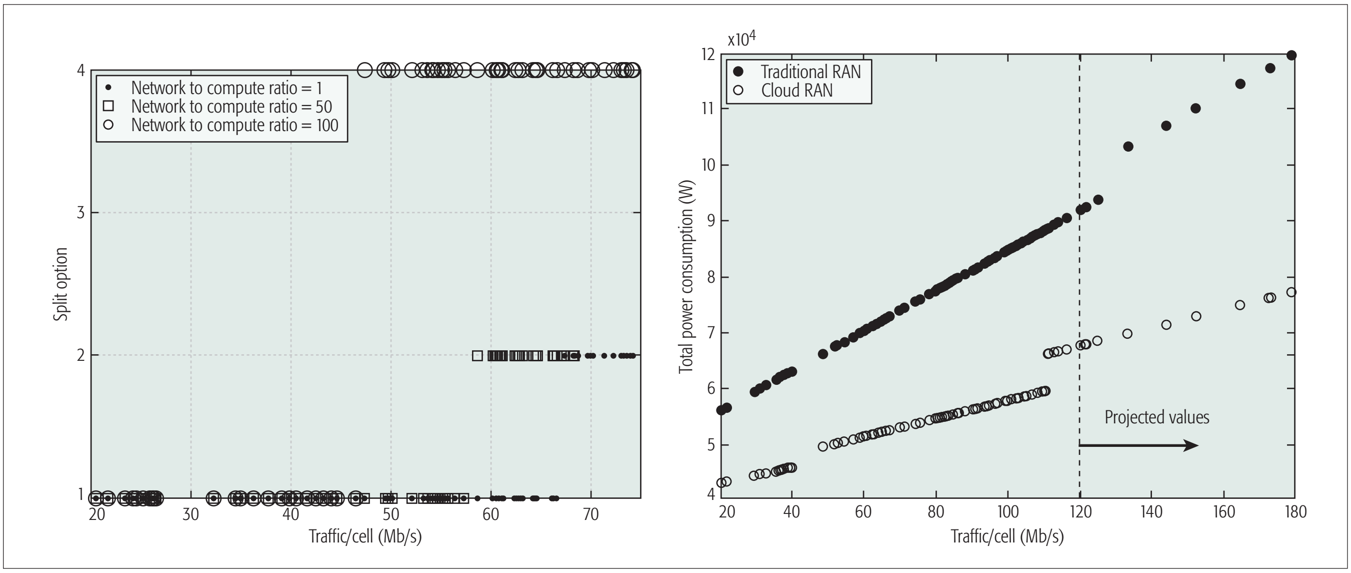


Figure 5. a) Split option as a function of load for different compute to network ratios; b) impact of the traffic load on the total power consumption.

Conclusions

This article presents a converged optical-wireless 5G infrastructure proposed by the EU 5GPPP project 5G-XHaul aiming to support jointly operational network and end-user services. The overarching architecture proposed is aligned with the SDN reference architecture and the ETSI NFV standard. A novel MOP modeling framework has been developed to evaluate the performance of the 5G-XHaul architecture taking into consideration the joint support of FH and BH services. Our modeling results show that the proposed architecture can offer significant benefits in terms of energy consumption but at the expense of end-user service delays.

Acknowledgment

This work has been supported by the EU Horizon 2020 5GPPP project 5G-XHaul.

References

[1] "C-RAN. The Road Towards Green RAN," White Paper 3.0, Dec. 2013; http://goo.gl/Sw6bfw

[2] Nokia Press Release, "Nokia Accelerates Centralized RAN Deployment with Expanded Mobile Fronthaul Solution #MWC16," Feb. 2016; http://tinyurl.com/h5kmbvb.

[3] U. Dötsch *et al.*, "Quantitative Analysis of Split Base Station Processing and Determination of Advantageous Architectures for LTE," *Bell Labs Tech. J*, vol. 18, no. 1, May 2013, pp. 105–28.

[4] D. Wubben *et al.*, "Benefits and Impact of Cloud Computing on 5G Signal Processing: Flexible Centralization through Cloud-RAN," *IEEE Signal Process. Mag.*, vol. 31, no. 6, Nov. 2014, pp. 35–44.

[5] N. Nikaein *et al.*, "OpenAirInterface: A Flexible Platform for 5G Research," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 5, Oct. 2014.

[6] ETSI GS NFV-SWA 001 V1.1.1 (2014-12), "Network Functions Virtualisation (NFV);Virtual Network Functions Architecture," 2014.

[7] H2020 Project 5G-Xhaul, Deliv. 2.2, "System Architecture Definition," July 2016;, http://www.5g-xhaul-project.eu/download/5G-XHaul_D_22.pdf.

[8] Bo Han *et al.*, "Network Function Virtualization: Challenges and Opportunities for Innovations," *IEEE Commun. Mag.*, vol. 53, no. 2, Feb. 2015, pp. 90–97.

[9] B. R. Rofoee *et al.*, "Hardware Virtualized Flexible Network for Wireless-DataCenter (invited)", *IEEE/OSA J. Opt. Commun. Net.*, Mar. 2015, vol.3, pp A526–A536.

[10] A. Pascoletti and P. Serafini, "Scalarizing Vector Optimization Problems," *J Optimization Theory Appl*, vol. 42, 1984, pp. 499–524.

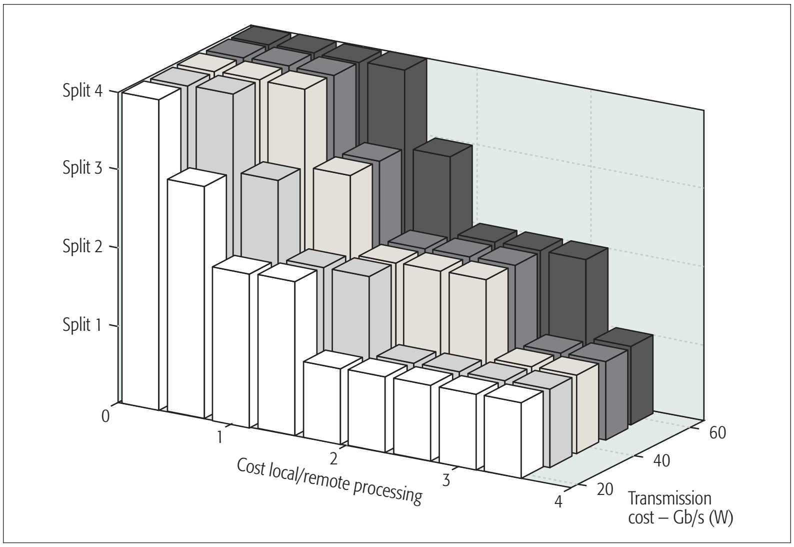


Figure 6. Split options for various processing and transmission costs.

[11] M. Savi, M. Tornatore, and G. Verticale, "Impact of Processing Costs on Service Chain Placement in Network Functions Virtualization," *Proc. IEEE NFV-SDN*, 2015, 18-21 Nov. 2015, pp. 191–97.

[12] J.-Y. Le Boudec and P.Thiran, *Network Calculus: A Theory of Deterministic Queuing Systems for the Internet*, Springer, 2001

[13] A. Tzanakaki *et al.*, "Virtualization of Heterogeneous Wireless-Optical Network and IT Infrastructures in Support of Cloud and Mobile Cloud Services," *IEEE Commun. Mag.* , vol. 51, no. 8, Aug. 2013, pp.155–61.

[14] J. Chang *et al.*, "Workload Diversity and Dynamics in Big Data Analytics: Implications to System Designers," *Proc. ASBD 2012*.

[15] X. Chen *et al.*, "Analyzing and Modeling Spatio-Temporal Dependence of Cellular Traffic at City Scale," *Proc. IEEE ICC*, 2015, pp.3585–91.

Biographies

Anna Tzanakaki (eanna.tzanakaki@bristol.ac.uk) is an assistant professor at the National and Kapodistrian University of Athens, Greece, and a research fellow at the University of Bristol, United Kingdom. She is a co-author of over 160 publications in international journals and conferences. Her research interests include converged networks, and network architectures, technologies, and protocols.

Markos Anastasopoulos holds a Diploma degree in electrical and computer engineering, an M.B.A. degree in financial engineering, and a Dr. Eng. degree from the National Technical

University of Athens. Currently, he is a researcher at the High Performance Networks Group of the University of Bristol working on the design of converged 5G infrastructures.

Ignacio Berberana's biography was not available.

Dimitris Syrivelis received his B.Sc. degree from the Technical University of Crete and his Ph.D. from the University of Thessaly. He was a postdoctoral fellow at the Centre for Research and Technology Hellas. He is currently a research engineer at the IBM Research Dublin lab and works on software defined infrastructures for hyperscale computing.

Paris Flegkas received a Diploma in electrical and computer engineering from Aristotle University, Greece, and an M.Sc. and a Ph.D. from the University of Surrey, United Kingdom. He is an adjunct lecturer and a senior researcher at the University of Thessaly. His research focuses on SDN, NFV management, and information-centric networking.

Thanasis Korakis is an assistant professor at the Department of Electrical and Computer Engineering at the University of Thessaly. He has published more than 60 papers and holds 5 patents. He was the PI for several NSF USA projects and has participated in several EU projects. He is the Associate Director of NITLab.

Daniel Camps-Mur is leading the Mobile and Wireless Internet group at I2CAT in Barcelona. Previously, he was a senior researcher at NEC Network Laboratories in Heidelberg, Germany. He holds a Master's degree and a Ph.D. from the Polytechnic University of Catalonia (UPC). His research interests include mobile networks and IoT.

Ilker Demirkol is a research professor in the Department of Network Engineering at UPC. His research focus is on communication protocol development and evaluation for wireless networks. He has also held several positions in industry as a network engineer, and a system and database consultant.

Jesús Gutiérrez Terán received his B.S. degree and Ph.D. in telecommunication engineering from the University of Cantabria in 2008 and 2013, respectively. Since 2013, he has been with IHP in Frankfurt (Oder), Germany. His research interests include digital signal processing for high-performance hardware architectures and millimeter-wave systems.

Eckhard Grass received his Dr.-Ing. degree in electronics from Humboldt University Berlin in 1993. After six years of research and lecturing in London, United Kingdom, since 1999 he has been with IHP, leading a research group on wireless broadband communications. Furthermore, he is a professor at Humboldt University Berlin since 2011.

Qing Wei received her M.Sc. degree in communication engineering from TU Munich, Germany, in 2001. From 2002 to 2015 she worked as a researcher/senior researcher at DOCOMO Euro Labs. In 2015, she moved to Huawei Technologies, working as a principal researcher in the area of 5G mobile network architecture and network programmability.

Emmanouil Pateromichelakis received his M.Sc. and Ph.D. degrees in mobile communications from the University of Surrey, United Kingdom, in 2009 and 2013 respectively. From 2013 to 2015 he was a postdoctoral fellow at 5GIC, University of Surrey. He is currently working as a senior researcher at Huawei Technologies, focusing on 5G and beyond solutions.

Nikola Vucic received his Dr.-Ing. degree from TU Berlin, Germany, in 2009. From 2003 to 2010, he was a research associate at Fraunhofer HHI, Berlin. Since 2011, he has been with Huawei Technologies in Munich, Germany, working as a senior researcher in the areas of wireless networks and future internet.

Albrecht Fehske received his Ph.D. from Vodafone Chair, TU Dresden in 2014 with highest honors. He co-authored more than 40 research publications. In 2013, he co-founded Airrays, a startup company, which delivers fully adaptive antenna technology for 4G and upcoming 5G radio access networks.

Michael Grieger received his Ph.D. from Vodafone Chair, TU Dresden in 2014. He has co-authored 36 research publications and is an inventor of 4 patents. Today, he is with Airrays GmbH, which he co-founded in 2013. Airrays delivers fully adaptive antenna technology for 4G and upcoming 5G radio access networks.

Michael Eiselt started his career in optical communications in 1989 and has worked for various companies and research organizations in Germany and the United States. As a director of Advanced Technology at ADVA Optical Networking, Germany, he is currently leading physical layer research for high-speed long-haul, data center interconnect, and access applications.

Jens Bartelt received his Dipl.-Ing. (M.S.E.E.) from Technische Universität Dresden, Germany. He is a research associate at the Vodafone Chair Mobile Communications Systems at TU Dresden, Germany, working toward his Ph.D. He is involved in several 5G research activities, including the EU projects iJOIN, 5G-XHaul, and the 5G Lab Germany at TU Dresden.

Gerhard P. Fettweis received his Dipl.-Ing. and Ph.D. degrees from Aachen University of Technology, Germany. Since September 1994 he has held the Vodafone Chair at Technische Universität Dresden, Germany. In 2012, he received an Honorary Doctorate from Tampere University. He is a well-known entrepreneur who has co-founded 13 start-ups and coordinates 2 DFG centers at TU Dresden

George Lyberopoulos has been involved in more than 30 EU and national research projects. He joined COSMOTE in 1999, and today he heads the Research & Development Department, Fixed and Mobile. He is an author of over 50 scientific papers in the areas of mobile telecommunications.

Eleni Theodoropoulou, MSc., is a telecom engineer. She has been working for Greek mobile operators since 1994. She has participated in several EU and national research projects and authored several scientific papers in the area of mobile communications. Since 2009, she has headed the R&D Projects Mobile Section of COSMOTE.

Dimitra Simeonidou is a professor at the University of Bristol, the Smart Internet lab director, chief scientific officer of Bristol Is Open, head of the High Performance Networks group and a Royal Society Wolfson scholar. She is a co-author of over 400 publications and 12 patents. Her research focuses on high-performance networks, SDN, and smart city infrastructures.

IEEE ICC™ 2018

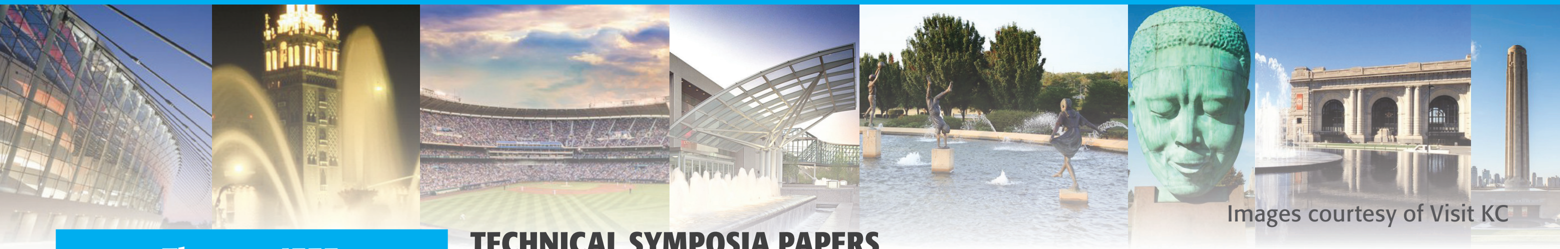
IEEE International Conference on Communications

Communications for Connecting Humanity

20-24 May 2018
Kansas City,
Missouri, USA

Back in the US after 15 years!

CALL FOR TECHNICAL & INDUSTRY SUBMISSIONS



Images courtesy of Visit KC

The 2018 IEEE International Conference on Communications (ICC) will include a Technical Program comprised of 13 specific symposia, tutorials and workshops as well as an Industry Program featuring panels, demonstrations, tutorials and workshops.

TECHNICAL SYMPOSIA PAPERS

Authors are invited to submit original technical papers in the following areas:

- Selected Areas in Communications
  - Access Systems and Networks
  - Big Data
  - Cloud Communications and Networks
  - Data Storage
  - E-Health
  - Internet of Things
  - Molecular, Biological and Multi-scale Communications
  - Smart Grid Communications
  - Powerline Communications
  - Social Networks
  - Satellite and Space Communications
  - Smart Cities
- Ad Hoc and Sensor Networking
- Cognitive Radio and Networking
- Communications and Information System Security
- Communications QoS, Reliability and Modelling
- Communications Software and Services
- Communication Theory
- Green Communications
- Next Generation Networking and Internet
- Optical Networks and Systems
- Signal Processing for Communications
- Wireless Communications
- Wireless Networking

IF&E Proposals

Proposals are sought that focus on latest topics, products and innovations of particular interest to industry and government in communications and networking.

Industry Demonstrations

Hardware and/or software demonstrations are sought that are meant to showcase new and innovative technology.

IMPORTANT DATES

Technical Symposia Papers
Due 15 October 2017

IF&E Proposals
Due 10 November 2017

Industry Demonstrations
Due 5 January 2018

Organizing Committee

General Chairs
Andrzej Jajszczyk,
AGH University of Science and Technology, Poland
Ron Marquardt, Sprint, USA
Executive Chair
Deep Medhi, University of Missouri-Kansas City, USA
Executive Vice Chair
Victor Frost, University of Kansas, USA
TPC Chair
Yi Qian, University of Nebraska-Lincoln, USA
TPC Vice Chairs
Rose Qiangyang Hu, Utah State University, USA
Lisandro Zambenedetti Granville,
Federal University of Rio Grande de Sul, Brazil
Workshop Co-Chairs
Bala Natarajan, Kansas State University, USA
Byrav Ramamurtny,
University of Nebraska-Lincoln, USA
Massimo Tornatore, Politecnico di Milano, Italy
Tutorials Co-Chairs
Tricha Anjali,
International Institute of Information Technology, Bangalore, India
Caterina Scoglio, Kansas State University, USA
Rosa Zheng,
Missouri University of Science & Technology, USA
IF&E Chair
Durga Satapathy, Sprint, USA

For more information, visit
http://icc2018.ieee-icc.org

IEEE

IEEE ComSoc™
IEEE Communications Society